Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea in Fisica

Caratterizzazione della misura di Single Node Entropy nell'ambito della Teoria Statistica dei Network

Relatore:
Prof. Daniel Remondini

Presentata da:
Lorenzo Posani

Correlatore:
Prof. Gastone Castellani
Dott.ssa Giulia Menichetti

Sessione II

Anno Accademico 2012/2013

ai miei genitori,
radici e linfa dei miei rami
e Federica,
sole delle mie foglie

Sommario

In questo lavoro si è affrontata la definizione e la caratterizzazione di una misura di entropia di singolo nodo nell'ambito della teoria statistica dei network, per ottenere informazioni a livello di singolo nodo a fini di analisi e classificazione. Sono state introdotte e studiate alcune proprietà di questi osservabili in quanto la Network Entropy, precedentemente definita e utilizzata nello stesso contesto, fornisce un'informazione globale a livello dell'intero network.

I risultati delle analisi svolte con questa definizione sono stati confrontati con una seconda definizione di entropia di singolo nodo proveniente dalla letteratura, applicando entrambe le misure allo stesso problema di caratterizzazione di due classi di nodi all'interno di un network.

Nel **capitolo 1** sono stati introdotti i fondamenti della meccanica statistica, in particolare la definizione di micro/macro stato nel contesto della teoria di Boltzmann, e le definizioni della teoria dell'informazione, con alcune applicazioni.

Tra questi due contesti è stato portato avanti un parallelismo volto a mettere in risalto il concetto di *Entropia*, definita nella fisica attraverso la costruzione teorica della meccanica statistica e i principi termodinamici, e nella teoria dell'informazione come contenuto medio informativo di una distribuzione di probabilità. In particolare si è visto che per alcune situazioni, come nella definizione della formula di Boltzmann $S = k \log W$, la teoria dell'informazione fornisce uno strumento di analisi in grado di portare allo stesso risultato della meccanica statistica.

Nel **capitolo 2** sono state inizialmente introdotte le principali definizioni della teoria matematica dei grafi. Sono stati quindi affrontati i due approcci di modellizzazione di un ensemble di network, *microcanonico* e *canonico*, approfondendone i metodi di imposizione dei vincoli e le differenze relative nelle assunzioni fondamentali.

È stata quindi data, in entrambi gli approcci, una definizione di Entropia nell'ambito della teoria dei network, seguendo i lavori di M. Newman, G. Bianconi e altri, attraverso gli strumenti della meccanica statistica e della teoria dell'informazione. L'entropia del network è infatti definita, secondo il formalismo microcanonico, come il logaritmo del numero di grafi che soddisfano certi vincoli specifici, e, nel formalismo canonico, come l'entropia di Shannon della matrice di probabilità descrittiva dell'ensemble di provenienza del network, trovata mediante la massimizzazione dell'entropia secondo i metodi propri della meccanica statistica. Sono stati quindi mostrati i calcoli relativi all'imposizione di diversi vincoli (random network, configuration model, spatial network) all'ensemble di network in entrambi i formalismi.

È stata poi definita una nuova misura, chiamata **Entropia di Singolo Nodo**, come l'entropia di Shannon del vettore relativo al singolo nodo nella matrice dell'ensemble.

Si è ripresa infine una seconda definizione di entropia su un singolo nodo, dai lavori di A. Teschendorff e S. Severini.

Nel **capitolo 3** Sono state caratterizzate queste due misure al fine di ottenerne un confronto critico. Inizialmente è stata fatta un'analisi matematica della misura qui definita a livello di distribuzione nel configuration ensemble, in cui è possibile utilizzare un approccio analitico, assumendo valide alcune approssimazioni. Sono stati poi implementati dei toy model, simulazioni di network, al fine di analizzare il comportamento delle due misure in diverse situazioni nel contesto degli *spatial network*, dove ai vincoli sulle connettività sono aggiunti vincoli secondo una matrice di distanze relative tra i nodi in una struttura geometrica. Obiettivo della caratterizzazione è valutare la performance delle misure nel distinguere gruppi di nodi con diverse caratteristiche. Ad esempio, se applicata a dati di trascrittomica (misure espressione genica immerse in un network di interazione di proteine), l'obiettivo è trovare un metodo per distinguere i comportamenti di singolo gene/proteina in diversi ambiti (tumore-sano, giovane-anziano..), sfruttando informazioni a livello dell'intero sistema (il network di interazione tra proteine).

Nel **capitolo 4** sono state riportate le conclusioni della caratterizzazione vista nel capitolo 3.

Indice

1 Teoria dell'Informazione e Meccanica Statistica **9**

1.1 Fondamenti di Teoria dell'informazione 9

1.2 Entropia H nella teoria dell'informazione 10

1.2.1 Definizione assiomatica 10

1.2.2 Definizione a posteriori 11

1.2.3 Compressibilità di una stringa 13

1.3 Elementi di Meccanica Statistica 14

1.3.1 Principio di Equiprobabilità e Ipotesi Ergodica 15

1.3.2 Metodo di Boltzmann 15

1.4 Entropia S in termodinamica 17

1.4.1 Principi sperimentali 17

1.4.2 Principio di Boltzmann 18

1.4.3 Entropia di Boltzmann vs. Entropia di Shannon 19

2 Entropia nei Network **21**

2.1 Definizioni nella Teoria dei Grafi 21

2.2 Rappresentazione di un Grafo 23

2.2.1 Matrice delle adiacenze 23

2.2.2 Vettore di connettività 23

2.3 Modellizzazione di un network 23

2.3.1 Ensemble statistico di network 23

2.3.2 Approccio microcanonico 24

2.3.3 Approccio canonico 24

2.4 Network Entropy nell'approccio microcanonico 25

2.4.1 Imposizione dei vincoli 25

2.4.2 Random graph 26

2.4.3 Configuration ensemble 26

2.4.4 Spatial ensemble 27

2.5 Network Entropy nell'approccio canonico 29

2.5.1 Random Network 30

2.5.2 Configuration Ensemble 30

2.5.3 Spatial Ensemble 31

2.6 Single Node Entropy S_i . . . 33
2.6.1 Definizione . . . 33
2.6.2 Definizione alternativa S_i^{TS} . . . 33

3 Caratterizzazione delle misure di Single Node Entropy 35
3.1 S_i nel Configuration Ensemble . . . 35
3.1.1 distribuzione delle S_i all'interno dell'ensemble . . . 35
3.1.2 Sensibilità alla connettività media $\langle k_i \rangle$ . . . 37
3.2 Spatial Ensemble . . . 40
3.2.1 Comportamento *entropico* . . . 41
3.2.2 Toy models per la Single Node Entropy . . . 42
3.2.3 μ model . . . 45
3.2.4 σ model . . . 47
3.2.5 Toy models per l'entropia di singolo nodo S^{TS} . . . 53
3.2.6 μ-model S^{TS} . . . 53
3.2.7 σ-model S^{TS} . . . 56

4 Conclusioni 59

Bibliography 63

Elenco delle figure

1 rappresentazione grafica del network di Internet, i nodi rappresentano gli indirizzi IP 7

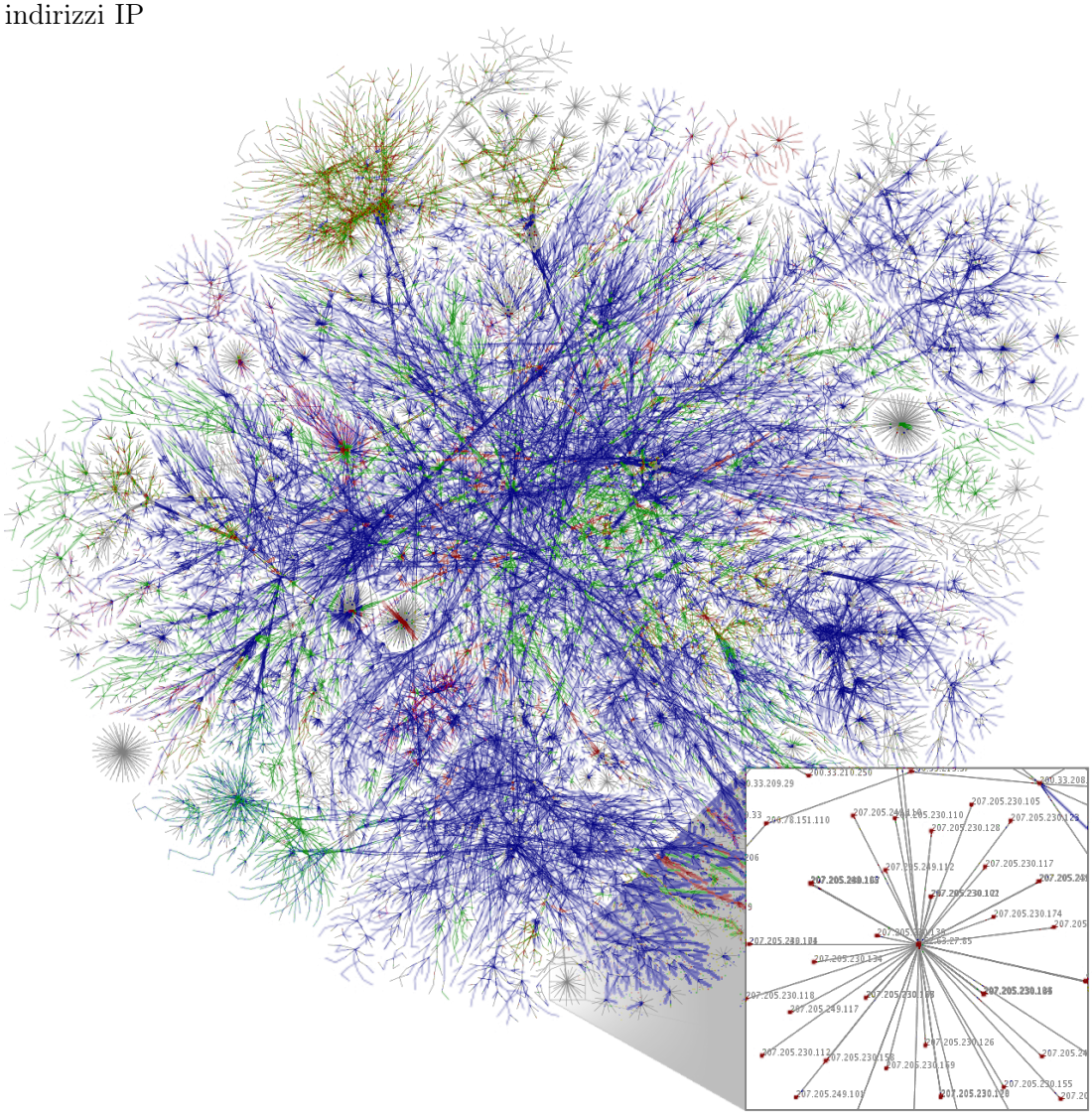
1.1 Struttura del Canale di trasmissione secondo Shannon 9
1.2 Entropia di Shannon nel caso binario: $H(p) = -p\log(p) - (1-p)\log(1-p)$ 12

2.1 rappresentazione grafica del network di interazione proteica nel lievito . 22
2.2 Rappresentazione dei network mediante matrice di adiacenza a_{ij} e vettore connettività $\vec{k}$ 24

3.1 Plot di S_i in funzione di k; $N = 1000$, $p_{link} = .1$ 36
3.2 Distribuzione delle entropie di singolo nodo S_i all'interno dell'ensemble configurazionale; $N = 10000$ 37
3.3 Distribuzione delle connettività k_i 37
3.4 Media delle S_{SN} al variare della p_{link} (in alto) e deviazione standard relativa (in basso) 38
3.5 Valori di media delle entropie di singolo nodo in funzione del logaritmo di N, con il relativo fit lineare. 40
3.6 Comportamento della media delle entropie di singolo nodo S_i^{TS} al variare della randomness nella matrice di distanza 42
3.7 Comportamento della media delle entropie di singolo nodo S_i al variare della randomness nella matrice di distanza 43
3.8 Valori di S_i nella situazione spaziale a riposo con $\mu_0 = 1$ e $\sigma_0 = 1$. In rosso i nodi appartenenti alla prima metà, in giallo quelli nella seconda metà. Non si osserva chiaramente differenziazione tra i due gruppi. . . . 44
3.9 Zoom della distribuzione 44
3.10 Valori di S_i nella situazione spaziale variata μ-**model** con $\mu_{var} = 100 \cdot \mu_0$ e $\sigma_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_μ 45
3.11 Valori di ΔS^0 tra la situazione spaziale variata μ-**model** con $\mu_{var} = 100$ e la situazione a riposo. 46
3.12 Zoom sulla distribuzione 46

3.13 Valori di S_i nella situazione spaziale variata σ-**model** con $\sigma_{var} = 100 \times \sigma_0$ e $\mu_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_σ 49
3.14 uno zoom sulla distribuzione mostra una separazione sensibile dei due sotto-gruppi 49
3.15 Valori di S_i nella situazione spaziale variata σ-**model** con $\sigma_{var} = 100 \times \sigma_0$ e $\mu_0 = 1$ in funzione dell'indice del nodo i nella matrice di adiacenza. . . 50
3.16 uno zoom sulla distribuzione mostra una separazione netta tra le due metà dei nodi, governate da diverse distribuzioni spaziali. 50
3.17 Valori di ΔS^0 tra la situazione spaziale variata σ-**model** con $\sigma_{var} = 100$ e la situazione a riposo. 51
3.18 Zoom sulla distribuzione 51
3.19 Valori di ΔS^{00} tra la situazione spaziale variata σ-**model** con $\sigma_{var} = 100$ e la situazione configurazionale, in funzione della connettività k_i. Non è mostrata dipendenza dalle connettività. 52
3.20 Distribuzione di ΔS^{00} all'interno del network. 52
3.21 Valori di S^{TS} nella situazione spaziale a riposo con $\mu_0 = 1$ e $\sigma_0 = 1$. In rosso i nodi appartenenti alla prima metà, in giallo quelli nella seconda metà. Non si osserva chiaramente differenziazione tra i due gruppi. . . . 54
3.22 Valori di media deviazione standard di S^{TS} per la *0-situation*, ossia tutti i nodi appartenenti alla stessa distribuzione 54
3.23 Valori di S^{TS} nella situazione spaziale variata μ-**model** con $\mu_{var} = 100 \times \mu_0$ e $\sigma_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_μ 55
3.24 Valori di S^{TS} nella situazione spaziale variata μ-**model** con $\mu_{var} = 100 \times \mu_0$ e $\sigma_0 = 1$, distribuiti in funzione dell'indice del nodo i 55
3.25 Valori di S^{TS} nella situazione spaziale variata σ-**model** con $\sigma_{var} = 100 \cdot \sigma_0$ e $\mu_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_σ 57
3.26 Valori di S^{TS} nella situazione spaziale variata σ-**model** con $\sigma_{var} = 100 \times \sigma_0$ e $\mu_0 = 1$ in funzione dell'indice del nodo i nella matrice di adiacenza 57

Figura 1: rappresentazione grafica del network di Internet, i nodi rappresentano gli indirizzi IP



Capitolo 1

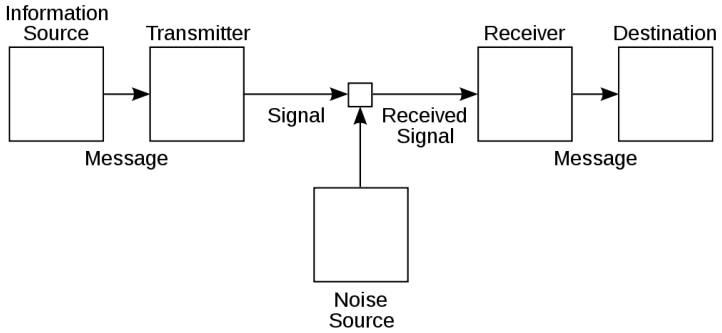
Teoria dell'Informazione e Meccanica Statistica

1.1 Fondamenti di Teoria dell'informazione

La nascita della teoria dell'informazione viene fatta risalire al 1948, anno in cui Claude Shannon, ritenuto il padre fondatore della teoria, pubblica un articolo dal titolo *A Mathematical Theory of Communication*, in cui poggia le basi per una teoria analitica della trasmissione e compressione di una quantità nota di dati [8].

In questo lavoro vengono definiti i concetti di *canale di trasmissione*, *sorgente*, *ricevitore* (vedi figura 1.1) e studiata la propagazione dell'informazione attraverso un canale con struttura e rumore noti. Viene mostrato quindi che esiste un limite superiore alla quantità di informazione che può essere comunicata ad ogni trasmissione, definita *Capacità C* del canale.

Figura 1.1: Struttura del Canale di trasmissione secondo Shannon



Viene inoltre affrontato il parallelismo tra una sorgente discreta di simboli, ad esempio un generatore di parole, e un processo stocastico markoviano, mostrando come un'analisi sufficientemente approfondita della lingua inglese può portare a un generatore di frasi coerenti.

Shannon vuole quindi definire una misura che tenga conto del *contenuto informativo* o dell'incertezza che un tale processo produce, definendo così l'*Entropia H*.

Sebbene sembri un campo molto specifico, gli sviluppi di questa teoria hanno portato a risultati del tutto generali che hanno profonde implicazioni in un'ampia varietà di discipline matematiche e fisiche, quali la fisica statistica, la complessità algoritmica, la teoria della probabilità e molte altre.

Nel lavoro qui esposto si farà uso soprattutto della definizione di entropia dell'informazione, che sarà dunque approfondita.

1.2 Entropia H nella teoria dell'informazione

Data una variabile aleatoria X che può assumere valori x all'interno di un alfabeto χ , definiamo $p(x)$ la probabilità che X assuma valore x.

L'entropia dell'informazione, o entropia di Shannon, di una variabile soggetta a tale distribuzione si definisce

$$H(X) = -\sum_{x \in \chi} p(x) \log(p(x)) \tag{1.1}$$

L'entropia di Shannon si occupa di misurare l'informazione a livello *simbolico*, ossia tenendo presente solo i vincoli a cui sottostanno i simboli e la loro mutua relazione, senza preoccuparsi del *significato* effettivo che questi assumono. Nella teoria dell'informazione si prende in esame una *sorgente* di simboli, che verrà qui descritta come una variabile aleatoria. In tal senso, *variabile aleatoria* e *sorgente* avranno qui un significato equivalente. Ci sono diversi modi per arrivare alla forma 1.1, partendo da presupposti logici o rispondendo a domande specifiche.

1.2.1 Definizione assiomatica

Il modo che Shannon utilizzò per formulare questa misura è chiamato *metodo assiomatico* e consiste nel definire una serie di assiomi che una misura di contenuto informativo di una sorgente deve soddisfare, per poi mostrare che la forma possibile di una tale funzione della distribuzione di probabilità è unica ed è della forma 1.1.

Ci sono inoltre diversi set di assiomi da cui si ricava la stessa definizione. Quelli qui scelti sono relativi alla misura di informazione di un singolo *simbolo*, data una variabile aleatoria sottostante una distribuzione di probabilità nota sui valori di un alfabeto χ con probabilità $p(x) = prob(X = x \in \chi)$. In tal senso, l'entropia di Shannon sarà definita come il valore atteso del contenuto informativo della sorgente di simboli. Il seguente set di assiomi non è precisamente lo stesso utilizzato da Shannon nel suo articolo, ma una rielaborazione intuitiva che porta agli stessi risultati.

Il *contenuto informativo* $i(E)$ di un evento, secondo Shannon, ha di fatto una relazione stretta con l'incertezza associata a questo evento. L'idea di fondo è che più l'osservatore sarà *sorpreso* nel vedere questo evento, più sarà alto il suo contenuto informativo, e viceversa. L'esempio più banale, ma significativo, è un evento con probabilità

certa: sicuramente non è in grado di contenere alcun tipo di informazione. Le proprietà assiomatiche di $i(x)$, $x \in \chi$, saranno quindi:

- $i(x)$ deve essere funzione continua della probabilità del simbolo $= F(p(x))$. Sarà quindi funzione definita da un intervallo $]0, 1]$ ai valori positivi (il contenuto informativo è una quantità positiva), escludendo lo 0 dato che un evento che non si verifica mai non è di nessun interesse e nel valore atteso verrebbe comunque eliminato dal limite $x \log(x) \to 0$ per $x \to 0$
- $F(1) = 0$, come detto prima, l'evento certo non porta informazione
- $F(p \to 0) \to +\infty$
- F deve essere funzione continua e monotona, dato che esprime una proprietà concettuale univoca e monotona funzione della probabilità.
- F deve seguire una composizione del tipo $F(a \cdot b) = F(a) + F(b)$, essendo la probabilità di due eventi indipendenti il prodotto delle singole probabiltà, ed essendo il contenuto informativo di una serie di eventi congiunti la somma delle specifiche informazioni. È intuitivo pensare a questa proprietà nel caso di un evento b che si verifica con un evento certo a: il contenuto informativo deve essere quello di b, quindi è naturale pensare alla somma $(i(b) + 0)$ come composizione delle informazioni.

Shannon dimostra[1] che la funzione $-\log(x)$ soddisfa tutte le proprietà ed è unica. L'entropia H viene quindi definita come il contenuto informativo medio di una sorgente, attraverso il valore atteso:

$$H(X) = \sum_{x \in \chi} p(x) i(x) = -\sum_{x \in \chi} p(x) \log(p(x)) \tag{1.2}$$

Un'idea dell'andamento della funzione $H(p)$ si può avere graficandola nel caso di un alfabeto binario, ossia nel caso in cui la distribuzione sia governata da una sola variabile $p = p_1$, con $p_2 = 1 - p_1$. In figura 1.2 è mostrato questo andamento: la funzione è simmetrica e si annulla per $p = 0, p = 1$ casi in cui non c'è incertezza o informazione, ed è massima nel caso $p = \frac{1}{2}$

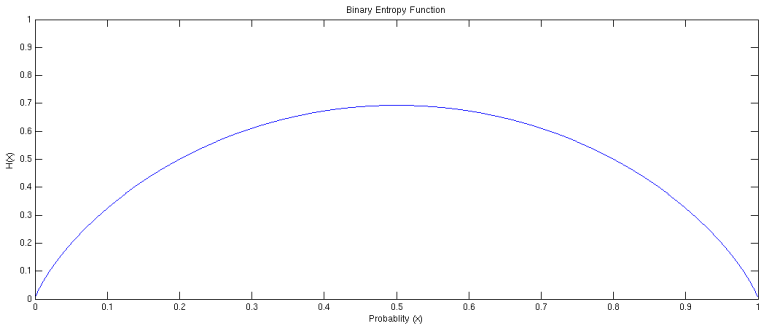
1.2.2 Definizione a posteriori

È possibile anche definire l'entropia H a partire da domande specifiche riguardo una distribuzione di simboli, a cui vogliamo che la misura risponda. In particolare poniamo una questione sulla quantità di informazione minima necessaria a specificare, in media, un'istanza particolare della distribuzione. Questa quantità definisce ottimamente il contenuto informativo di una sorgente, in quanto è direttamente riconducibile a una definizione operativa dell'entropia H [2].

[1] Per la dimostrazione dettagliata della forma di H si veda l'appendice 2 di [8].

[2] Questo approccio *operativo* è quello proposto in [10]

Figura 1.2: Entropia di Shannon nel caso binario: $H(p) = -p\log(p) - (1-p)\log(1-p)$



Pensiamo infatti ad un alfabeto di N simboli $\chi = x_1, x_2, ..., x_N$ con probabilità diverse $p_1, p_2, ..., p_N$ di essere prodotti da una sorgente.

Se volessimo comunicare una serie di simboli prodotti, potremmo etichettarli tutti con un indice numerico da 0 a $N-1$ e comunicare una successione di questi indici: avremo quindi bisogno di $\log_2(N)$ bit per ogni simbolo, e l'informazione media necessaria a trasmettere una serie di eventi sarà appunto $\log_2(N)$ bit.

Tuttavia, conoscendo a priori la distribuzione di probabilità a cui sottosta la sorgente, possiamo etichettare con un numero più basso di bit i simboli più probabili, e salire differenziando sulla dimensione dell'etichetta man mano che la probabilità diventa inferiore. La media della quantità di informazione necessaria a trasmettere una serie di simboli sarà quindi in genere inferiore alla semplice etichetta numerica, e di fatto sarà il valore atteso della grandezza dell'etichetta. Utilizzando come grandezze l'inverso della probabilità dell'evento, si ha esattamente la definizione di Shannon per l'entropia H:

$$H(X) = \sum_{x\in\chi} p(x)\log_2(\frac{1}{p(x)}) = -\sum_{x\in\chi} p(x)\log_2(p(x)) \tag{1.3}$$

Da questo punto di vista l'entropia di Shannon è definita come la *quantità minima di informazione necessaria in media per descrivere una serie di istanze della sorgente.* Con il logaritmo in base 2 si misura in *bit*, con il logaritmo naturale l'unità è il *nat*

esempio Prendiamo un alfabeto di 8 elementi con probabilità

$$p_i = [\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}, \frac{1}{32}]$$

Utilizzando un'etichettatura numerica avremo bisogno di $\log_2(8) = 3$ bit per comunicare ogni simbolo generato dalla sorgente, quindi la comunicazione costerà in media, appunto, 3 bit.

Se assegniamo invece un'etichetta di lunghezza proporzionale a $l_i = -\log_2(p_i)$, del tipo

$$0, 10, 110, 1110, 111100, 111101, 111110, 111111$$

la lunghezza media della comunicazione sarà uguale a $-\sum_i p_i \log_2(p_i) = 2$ bit. Abbiamo quindi risparmiato, in media, un bit per ogni simbolo comunicato.

1.2.3 Compressibilità di una stringa

Una delle caratteristiche principali dell'entropia di Shannon è che rappresenta il limite di compressibilità di una stringa di simboli. Questa intepretazione è inuitiva a partire da quella di *contenuto informativo*: è evidente che, a meno di perdere l'informazione contenuta nella stringa, esiste una lunghezza minima che ne rappresenta il contenuto. Per meglio capire il metodo di compressione, analizziamo il caso binario, $\chi = [0, 1]$, che rappresenta l'ambiente delle stringhe nei computer.

Data una stringa di lunghezza N, possiamo schematizzarla come un'istanza particolare di un *ensemble*[3]. di stringhe che hanno la stessa quantità di 1 e di 0. Questo pool è codificato quindi come la famiglia di tutte le stringhe di N cifre, con probabilità $p = prob(x = 1)$, e di conseguenza $prob(x = 0) = 1 - p$, determinata .

La quantità di informazione minima necessaria per riprodurre la stringa è quindi un'etichetta che codifica l'istanza particolare all'interno della famiglia, oltre al numero N e alla probabilità p che identificano l'ensemble in questione. Per calcolare l'ampiezza di questa etichetta, è sufficiente mostrare quante stringhe diverse possono essere create a parità di N e p, ossia la popolazione della famiglia relativa.

Le diverse possibili stringhe appartenenti alla stessa famiglia sono tutte le combinazioni di N elementi presi Np alla volta, ossia

$$n_{pool}(N, p) = \binom{N}{Np} = \frac{N!}{(N - Np)!(Np)!} = \frac{N!}{n_0! n_1!} \tag{1.4}$$

Se vogliamo identificare un'istanza particolare all'interno di questo ensemble attraverso un numero espresso in binario, la lunghezza S, in bit, dell'etichetta necessaria sarà il logaritmo in base 2 del numero di oggetti in esso contenuti.

Da qui è possibile calcolare un valore approssimato del coefficiente di compressibilità C, ossia il rapporto tra la lunghezza in bit minima per la codifica della stringa e il numero di bit di partenza N, in funzione dell'entropia H, utilizzando la formula di Stirling:

[3]Questa interpretazione è un primo esempio della fusione dei concetti di teoria dell'informazione con quelli propri della meccanica statistica, come l'ensemble statistico. Per un maggiore approfondimento del contesto si veda anche [12]

$$
\begin{aligned}
S(N,p) =& \log_2(n_{pool}(N,p)) \\
=& N\log_2(N) - N - n_0\log_2(n_0) + n_0 - n_1\log_2(n_1) + n_1 + O(\log(N)) \\
=& N\log_2(N) - n_0\log_2(n_0) - n_1\log_2(n_1) + O(\log(N)) \\
=& N\log_2(N) - n_0\log_2(N) - n_0\log_2(p_0) - n_1\log_2(N) - n_1\log_2(p_1) + O(\log(N)) \\
=& -Np\log_2(p) - N(1-p)\log_2(1-p) + O(\log(N)) \\
\implies C =& \frac{S(N,p)}{N} \simeq -p\log_2(p) - (1-p)\log_2(1-p) \\
=& -\sum_{x\in[0,1]} p(x)\log_2(p(x)) = H(p)
\end{aligned}
$$

A cui va aggiunto il numero di bit necessari a specificare il pool, ossia un ordine di $\log_2 N$, dello stesso ordine di approssimazione della formula di Stirling. Si trova dunque:

$$S(N,p) = NH(p) + O(\log_2(N)) \tag{1.5}$$

Che per grandi valori di N cresce linearmente con coefficiente H.

Si è trovato quindi che l'entropia H rappresenta il limite teorico di compressibilità per stringhe molto grandi, a meno di un ordine di $\frac{\log(N)}{N}$ che appunto si annulla a infinito.

1.3 Elementi di Meccanica Statistica

Ai fini di utilizzare l'approccio statistico nella teoria dei network e di definire il concetto di entropia in questo contesto, è utile rivedere alcuni concetti di base della meccanica statistica, che saranno successivamente ripresi nel formalismo dei network.

Questo approccio è stato originariamente utilizzato da Boltzmann alla fine del diciannovesimo secolo per mostrare come alcune proprietà macroscopiche della materia, come temperatura, pressione etc., emergano naturalmente dalle interazioni fondamentali tra i suoi costituenti (ad esempio, nel caso dei gas, le particelle) e dalla trattabilità statistica dell'insieme considerato, grazie al grande numero di costituenti in gioco.

Una rivoluzionaria implicazione di questa teoria è che la stabilità apparente di un sistema macroscopico è in realtà il risultato di un *equilibrio statistico*, in cui i costituenti, pur mantenendo singolarmente un comportamento sensibilmente dinamico, si trovano in uno stato globalmente ordinato.

È inoltre affascinante come una costruzione teorica inizialmente inventata come modello per la realtà fisica sia in realtà del tutto generale e, come vedremo, presenta una metodologia di analisi applicabile a campi anche profondamente diversi da quello delle particelle dinamiche.

1.3.1 Principio di Equiprobabilità e Ipotesi Ergodica

Il principio fondamentale della meccanica statistica è il cosiddetto principio di *equiprobabilità*, il quale asserisce che, dato un sistema hamiltoniano ad energia E nella sua descrizione nello spazio delle fasi, ogni punto appartenente alla superficie isoenergetica, chiamato *microstato*, ha uguale probabilità di essere occupato dal sistema.

Questo principio è di fatto una conseguenza dell'ipotesi *ergodica*, la quale asserisce che la traiettoria di un sistema che presenta hamiltoniana separabile ed è in grado di riorganizzare la distribuzione dei quanti di energia attraverso urti quasi istantanei, a lungo termine occuperà densamente la superficie isoenergetica nello spazio delle fasi compatibile con il suo stato dinamico, e che quindi, su tempi molto più lunghi dei tempi propri molecolari, è coerente ipotizzare che tutti i punti isoenergetici siano occupati con uguale probabilità.

Boltzmann utilizzò infatti questa proprietà per asserire, mediante il teorema di Liouville, che una quantità termodinamica, definita come una media integrale temporale di una proprietà del sistema sul percorso nello spazio delle fasi, può in realtà essere calcolata semplicemente come media *spaziale* nello spazio delle fasi.

Questo risultato è intuitivo se si pensa al principio di equiprobabilità applicato in molteplici istanti temporali successivi: se il sistema occupa tutto lo spazio occupabile, un integrale temporale diventerà di fatto un integrale sulla superficie equienergetica.

1.3.2 Metodo di Boltzmann

Prendiamo un sistema di N particelle identiche e distinguibili, per il quale asseriamo valga l'ipotesi ergodica. Etichettiamo con ϵ_α i valori di energia che ogni particella può occupare, ognuno avente una molteplicità interna g_α.

Questo sistema è descritto nello spazio delle fasi da un vettore $6N$-dimensionale, chiamiamolo $\vec{z}$. Boltzmann utilizza un cambio di variabile, ossia da $\vec{z}$ a $\vec{n}$, vettore nello *spazio delle popolazioni* del sistema. Con *popolazione* si intende una disposizione particolare delle particelle all'interno della struttura di livelli energetici, schematizzata dai numeri di occupazione n_α.

Troviamo quindi che lo spazio occupabile dal nostro vettore di popolazioni dovrà sottostare ai vincoli di conservazione dell'energia e del numero di particelle, dunque si ha che la superficie Γ esplorata ergodicamente dal sistema sarà

$$\Gamma = \{\vec{n} : \sum_\alpha n_\alpha \epsilon_\alpha = E, \sum_\alpha n_\alpha = N\} \tag{1.6}$$

Essendo questo cambio di variabile non iniettivo, e tenendo presente il principio di equiprobabilità dei punti nello spazio delle fasi, è naturale che i singoli vettori popolazione $\vec{n}$ non siano più equiprobabili. Per trovare le singole probabilità è tuttavia sufficiente vedere quanti stati originali diversi mappino nello stesso vettore connettività. Con un'adeguata normalizzazione, questo numero sarà equivalente alla probabilità che

il sistema si trovi nello stato descritto dal vettore $\vec{n}$ in esame. A questo punto la distribuzione della probabilità si riduce ad una semplice *conta* statistica dei diversi modi di disposizione interni che ogni vettore popolazione può presentare. Ciascuna disposizione interna è un *microstato*, mentre il vettore popolazione corrispondente è un esempio di *macrostato*.

I microstati relativi ad ogni vettore $\vec{n}$ sono dunque

$$\begin{aligned} W_{boltz} &= C_{N,n_0} D'_{g_0,n_0} \cdot C_{N-n_0,n_1} D'_{g_1,n_1} \cdot \dots \qquad (1.7) \\ &= \binom{N}{n_0} g_0^{n_0} \binom{N-n_0}{n_1} g_1^{n_1} \dots \\ &= \frac{N!}{n_0!(N-n_0)!} g_0^{n_0} \cdot \frac{(N-n_0)!}{n_1!(N-n_0-n_1)!} g_1^{n_1} \dots \\ &= N! \prod_{\alpha=0}^{\infty} \frac{g_\alpha^{n_\alpha}}{n_\alpha!} \qquad (1.8) \end{aligned}$$

La probabilità relativa a una singola configurazione diventa quindi

$$P(\vec{n}) = \frac{W_{boltz}(\vec{n})}{\sum_{\vec{n}_i \in \Gamma} W_{boltz}(\vec{n}_i)} \qquad (1.9)$$

Con l'approssimazione di Stirling la 1.8 può essere riscritta come

$$\log(W_{boltz}) = N \log N - N + \sum_\alpha n_\alpha \left[\log(g_\alpha) - \log(n_\alpha) - 1\right] \qquad (1.10)$$

Boltzmann dimostra inoltre che al crescere di N la distribuzione di probabilità tende ad essere sempre più piccata sul vettore $\vec{n}_{max}$ che *massimizza i microstati*, ossia presenta il numero massimo di stati che vi mappano. A tendere di N all'infinito, si ottiene una delta di Dirac su questa particolare disposizione, che risulta quindi essere l'unica possibile.

$$\lim_{N\to\infty} P(\vec{n}) = \delta(\vec{n} - \vec{n}_{max}) \qquad (1.11)$$

Il calcolo di una quantità termodinamica, definita precedentemente come la media sulla superficie isoenergetica di una certa funzione dello stato, si riduce quindi semplicemente al calcolo della funzione in esame, con i dovuti cambi di varabili, nella configurazione relativa al macrostato $\vec{n}_{max}$.

È qui che per la prima volta emerge l'ordine macroscopico da pure considerazioni statistiche e della struttura microscopica del sistema, e viene quindi definito il concetto di *equilibrio statistico*, profondamente differente dall'equilibrio dinamico.

Definita la quantità $W_{boltz}(\vec{n})$ che corrisponde al numero di microstati che mappano su $\vec{n}$, il massimo di questa funzione si trova con il metodo dei moltiplicatori di Lagrange, ossia risolvendo il sistema di equazioni in α

$$\frac{\partial \log W_{boltz}}{\partial n_\alpha} = \gamma + \beta n_\alpha \qquad (1.12)$$

Dove il moltiplicatore β si riferisce al vincolo di conservazione dell'energia e γ alla conservazione del numero di particelle.

Si trovano quindi le espressioni delle singole occupazioni, per un sistema all'equilibrio termodinamico, dei livelli energetici:

$$n_{\alpha\, max} = N\, \frac{g_\alpha e^{-\beta\epsilon_\alpha}}{Z_c} \tag{1.13}$$

Dove $Z_c = \sum_\alpha g_\alpha e^{-\beta\epsilon_\alpha}$ è detta *funzione di ripartizione canonica* e gioca un ruolo fondamentale nel calcolo di ogni proprietà del sistema nella termodinamica statistica.

1.4 Entropia S in termodinamica

Il concetto di Entropia nasce originariamente da considerazioni sperimentali: è noto fin dall'inizio del diciannovesimo secolo che esiste una funzione di stato correlata con le quantità termodinamiche (Calore Q, temperatura T), mediante una relazione ben precisa. Questa funzione, chiamata Entropia S, viene in questo contesto interpretata come *disordine* nel sistema.

Le relazioni tra le variabili termodinamiche antecedenti la costruzione di Boltzmann sono state espresse su basi puramente empiriche, effettuando misure sui gas e osservando come alcuni processi termodinamici siano possibili e altri no, senza un vero e proprio formalismo teorico in grado di giustificare le osservazioni.

Le variabili che definiscono lo *stato* termodinamico del sistema sono pressione P, volume V e temperatura T, ed uno stato è univocamente determinato da due di queste. Le relazioni che legano queste quantità con il concetto di *calore* come flusso di energia tra corpi con differente temperatura, e la stessa entropia S, sono note come i tre *principi della termodinamica*

1.4.1 Principi sperimentali

Principio 0 Il principio 0 della termodinamica definisce il concetto di *equilibrio termico* come la relazione tra due corpi che posti a contatto non mostrano scambio di calore, relazione sottostante a proprietà transitiva (se A è in equilibrio con B e B è in equilibrio con C, allora A è in equilibrio con C). In particolare definisce il rapporto tra equilibrio termico e Temperatura, che risulta essere la quantità condivisa dai corpi in questione.

I Principio

$$\delta Q = dE + \delta L \tag{1.14}$$

il primo principio è di fatto una rielaborazione della conservazione dell'energia, e stabilisce che gli scambi di calore altro non sono che scambi di energia.

II Principio

$$\delta Q_{rev} = TdS \tag{1.15}$$

Nel secondo principio viene introdotta la distinzione tra trasformazioni *reversibili*, le quali passano solamente tra stati di equilibrio termodinamico, e *irreversibili*. Questo principio asserisce che gli scambi di calore siano legati dalla relazione differenziale 1.15 ad una funzione di stato S, con proporzionalità data dalla temperatra T.

Viene completato nella forma generale:

$$\delta Q \leq TdS \tag{1.16}$$

1.4.2 Principio di Boltzmann

È ovviamente possibile fondere questi principi empirici con la costruzione teorica di Boltzmann, attraverso alcune considerazioni:

notiamo innanzitutto che l'incremento di lavoro δL può essere espresso come variazione delle variabili termodinamiche come $\delta L = PdV + VdP$. Utilizzando la descrizione dei livelli energetici nel metodo di Boltzmann, dalla 1.6 possiamo scrivere invece la relazione

$$dE = \sum_\alpha \epsilon_\alpha dn_\alpha + \sum_\alpha d\epsilon_\alpha n_\alpha = \delta Q - \delta L \tag{1.17}$$

Dove assumiamo che l'effetto della trasmissione di calore sia uno spostamento delle particelle tra i livelli energetici, e che il lavoro compiuto sul sistema risulti in una modifica dei valori di energia degli stati. Otteniamo quindi

$$\delta Q = \sum_\alpha \epsilon_\alpha dn_\alpha \implies \beta\delta Q = \sum_\alpha dn_\alpha(\beta\epsilon_\alpha + \gamma) \tag{1.18}$$

Dove l'aggiunta di γ è lecita nell'assunzione $\sum_\alpha dn_\alpha = 0$, ossia che si conservi il numero di particelle.

Dalla 1.12 con la 1.18 otteniamo

$$\beta\,\delta Q = \sum_\alpha dn_\alpha \frac{\partial \log W_{boltz}}{\partial n_\alpha} = d\,\log\,W_{boltz} \tag{1.19}$$

Il che porta alla ridefinizione di entropia nel formalismo di Boltzmann, si ha infatti con la 1.16 che

$$dS = \frac{1}{T}\,\delta Q = \frac{\beta}{T}\,d\,\log\,W_{boltz} \implies S = k_b\,\log\,W_{boltz} \tag{1.20}$$

la 1.20 è nota anche col nome di *Principio di Boltzmann* e si trova incisa nella nota forma $S = k\ \log\ W$ come epitaffio sulla tomba dello stesso Boltzmann a Vienna.

1.4.3 Entropia di Boltzmann vs. Entropia di Shannon

È curioso osservare come un approccio basato sulla teoria dell'informazione porti agli stessi risultati dell'approccio statistico di Boltzmann.

Poniamoci infatti in uno stato di equilibrio: ogni microstato tra i W possibili è occupabile dal sistema con probabilità uniforme, $p_i = \frac{1}{W}$. Applicando la misura di entropia di Shannon al sistema si ottiene

$$H = -\sum_{i=0}^{W} p_i \ \log(p_i) = -\sum_{i=0}^{W} \frac{1}{W} \ \log(\frac{1}{W}) = -\ \log(\frac{1}{W}) = \log(W) = \frac{1}{k_b}\ S_{boltz} \quad (1.21)$$

La 1.21 mostra che l'entropia di Shannon del sistema è identica alla formulazione dell'entropia di Boltzmann, a meno della costante k_b, che emerge per considerazioni relative alle unità di misura utilizzate in termodinamica classica.

Esiste inoltre un concetto più generale di entropia *fisica*, valido anche fuori dagli stati di equilibrio: la cosiddetta *entropia di Gibbs*

$$S_{gibbs} = -k_b \sum_i p_i \ \log(p_i) \quad (1.22)$$

Dove i è l'etichetta degli stati accessibili al sistema in esame. Si nota immediatamente la forma identica all'entropia di Shannon. È infatti da questa somiglianza che Shannon diede il nome Entropia alla sua definizione.

È curioso il fatto che Shannon non utilizzò mai la teoria di Boltzmann o di Gibbs per ricavare la sua misura di informazione, e chiamò questa misura Entropia solo poco prima della pubblicazione, grazie ad una discussione con Von Neumann. Shannon stesso avrebbe dichiarato:

> My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage. [11]

Queste analogie mostrano come le due definizioni di entropia riflettano in realtà un unico concetto fondamentale, che esplicita il ruolo dell'informazione nella descrizione dei sistemi fisici. [4].

[4] Questa generalizzazione non è chiaramente oggetto di studio nella presente tesi, la fusione dei due concetti in un'unica teoria generale è trattata ampiamente in letteratura, si veda ad esempio [5]

Capitolo 2

Entropia nei Network

In questo capitolo verranno presentate le definizioni fondamentali della teoria dei grafi, branca della matematica dalla quale deriva la definizione di Network che è stata utilizzata in questo lavoro.

I grafi sono di notevole interesse nei lavori contemporanei, in quanto rappresentano uno strumento di modellizzazione e studio per una grande varietà di realtà, quali le reti sociali, il traffico su rete stradale, Internet, le reti di distribuzione (elettrica, telefonica ecc.), le reti metaboliche in biologia teorica, o per le reti proteomiche e trascrittomiche nella System Biology.

Attraverso la simulazione dei network, infatti, si cerca di ricreare le condizioni che hanno portato ad un comportamento emergente globale simile a quello osservato nella realtà, per poi lavorare, ad esempio, su punti di controllo nevralgici (come nelle simulazioni del traffico) ai fini di ottenere una maggiore efficienza, o per cercare osservabili che permettano di distinguere diverse situazioni ad alto livello attraverso un'analisi algoritmica [6].

I grafi sono stati inoltre studiati in modo approfondito nella teoria matematica nella quale la loro definizione prende forma, già sviluppata dal celebre matematico Erdős [4] negli anni '60.

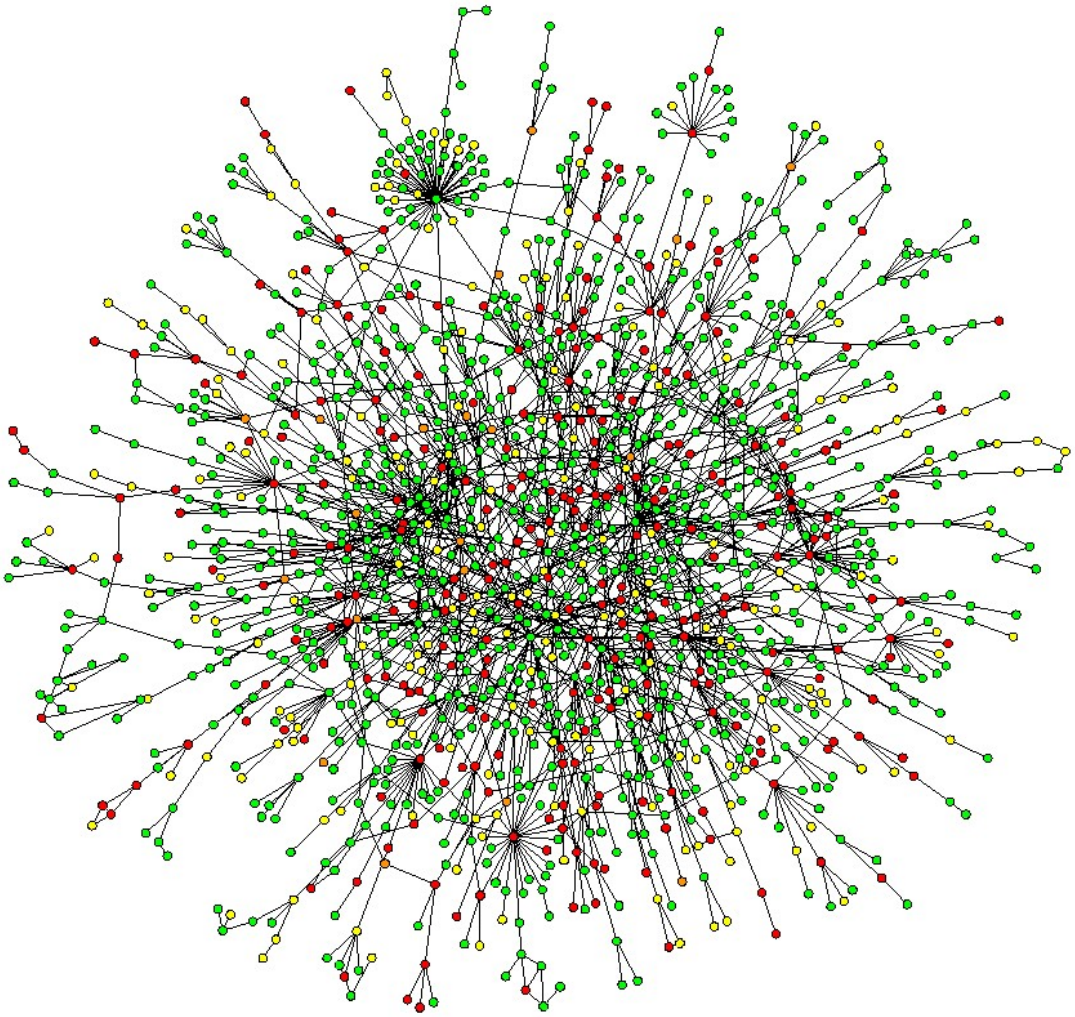
Verrà qui successivamente mostrato come sia possibile utilizzare gli strumenti della meccanica statistica per definirne una misura di *entropia* nell'ambito della teoria dei network, secondo la linea di lavoro di M.Newman, G. Bianconi e altri [2, 7].

2.1 Definizioni nella Teoria dei Grafi

Un grafo, o network, è un oggetto discreto costituito di *nodi*, o vertici, e di *connessioni* tra i nodi.

Def. Un grafo G è una coppia ordinata $G = (V, E)$ dove V è un insieme discreto di vertici, e E è un insieme di coppie (a, b) ; $a, b \in V$ che rappresentano le connessioni tra i vertici. Le connessioni possono essere *pesate* e/o *dirette*.

Figura 2.1: rappresentazione grafica del network di interazione proteica nel lievito



Def. Un grafo che presenta non più di una connessione tra una coppia di nodi è detto *semplice*.

Se sono permesse multiple connessioni tra una stessa coppia di nodi il grafo è detto *multigrafo*

Def. Un grafo è detto *completo* se presenta tutte le connessioni tra i vertici, ossia tutte le possibili $\binom{N}{2} = \frac{1}{2}N(N-1)$ coppie di nodi sono presenti in E.

Def. Un grafo si dice *connesso* se presenta una struttura topologica connessa, ossia esiste un cammino (successione di connessioni) per andare da qualunque nodo a qualunque nodo.

Def. Si definisce **connettività** k_i del nodo $i-$esimo il numero di primi vicini connessi al nodo i, ossia il numero di nodi a cui è connesso direttamente.

Come vedremo, il ruolo della connettività è centrale nello studio dei network, in particolare per la definizione dei vincoli nell'approccio entropico.

2.2 Rappresentazione di un Grafo

2.2.1 Matrice delle adiacenze

Def. La matrice delle adiacenze, o matrice delle connessioni, di un network a N nodi, è una matrice $N \times N$ definita come:

$$a_{ij} = \begin{cases} w_{ij}, & \text{se il nodo } i \text{ è connesso al nodo } j \\ 0, & \text{altrimenti} \end{cases} \tag{2.1}$$

dove w_{ij} è il peso relativo al link tra il nodo i e il nodo j.

Questa matrice costituisce di fatto l'informazione necessaria a identificare il network. Nel caso di connessioni semplici, non pesate, non dirette e senza auto-connessioni, la matrice sarà composta solo da valori 1 e 0, simmetrica e con diagonale nulla. Per semplicità nella successiva analisi saranno considerate solo matrici di questo tipo, senza tuttavia compromettere la generalità dei metodi utilizzati.

2.2.2 Vettore di connettività

Def. Il vettore delle connettività $\vec{k}$ è definito come il vettore delle somme sulle righe della matrice di adiacenze.

$$\vec{k} = (k_1,\ k_2,\ \ldots\ ,\ k_N) \quad \text{dove } k_i = \sum_j a_{ij} \tag{2.2}$$

Nel caso di network semplici e non pesati, ogni sua componente k_i è semplicemente il numero di primi vicini conessi al nodo i.

2.3 Modellizzazione di un network

I metodi di modellizzazione di un network possono essere molteplici, e sono fondati sul concetto di *ensemble*.

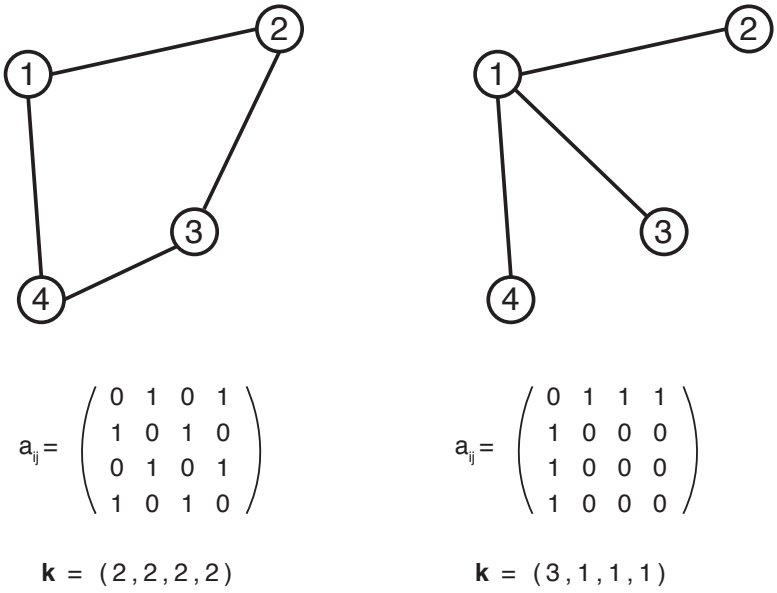
2.3.1 Ensemble statistico di network

Ciascun modello, allo stato attuale dei lavori [7], si concentra infatti non sullo specifico network, ma su un insieme di grafi che si presume possano *svolgere in egual modo* i compiti affidati al network specifico.

Si suppone infatti che il nostro grafo reale sia una rappresentazione particolare di una famiglia più generale di appartenenza, e si procede allo studio delle proprietà a livello di insieme.

Questo insieme è definito attraverso l'imposizione di particolari vincoli elaborati a partire da osservabili specifici estratti dal network reale. A seconda dell'imposizione di questi vincoli, in analogia alle definizioni in meccanica statistica, si distinguono gli approcci *microcanonico*, dove i vincoli sono soddisfatti da ogni network nell'ensemble, e *canonico*, dove i vincoli sono imposti solo come media sull'insieme.

Figura 2.2: Rappresentazione dei network mediante matrice di adiacenza a_{ij} e vettore connettività $\vec{k}$



Al fine di comprendere e modellizzare un network reale nel contesto degli ensemble statistici si fa un'assunzione di base:

Proposizione. *Il network reale costituisce un'istanza particolare di un ensemble di network, all'interno del quale costituisce la rappresentazione più probabile. [7]*

2.3.2 Approccio microcanonico

L'ensemble microcanonico è definito come l'insieme totale dei network che soddisfano in modo esatto una serie di vincoli dati.

L'approccio microcanonico, utilizzato da G. Bianconi [2, 3], si fonda sul concetto di *informazione* contenuta nei vincoli definenti l'ensemble di appartenenza del network reale in analisi.

Questo ensemble può essere costruito, a partire dal network reale, imponendo vincoli per gradi di approssimazione successivi (numero di link, gradi di connettività, etc.), diminuendo ogni volta il numero totale di grafi in grado di soddisfarli. Una trattazione più approfondita è data nella sezione 2.4

2.3.3 Approccio canonico

L'ensemble canonico è definito attraverso l'utilizzo e la massimizzazione dell'entropia di Gibbs dell'ensemble stesso, imponendo dei vincoli in modo più leggero rispetto all'approccio microcanonico:

si estraggono degli osservabili dal network reale, e si cerca quale sia l'insieme, o meglio la distribuzione di probabilità che governa l'insieme di network, che meglio modellizzi la provenienza del network specifico, imponendo che il valore atteso degli osservabili nell'ensemble sia uguale al valore esibito dal nostro network reale.

Per trovare quale distribuzione di probabilità meglio modellizzi questo ensemble si massimizza l'entropia di Gibbs della stessa distribuzione attraverso il metodo dei massimi vincolati di Lagrange, imponendo i vincoli come appena descritto.

L'imposizione dei vincoli è quindi più *soft* che nel caso canonico, dovendo i valori del network reale essere esibiti soltanto *in media* dall'ensemble trovato.

Le variabili che descrivono l'ensemble possono essere $P(G)$, probabilità dello specifico grafo, oppure p_{ij}, le probabilità globali del link tra il nodo i e il nodo j. L'entropia sarà quindi definita in modo diverso nei due casi, ma con opportune considerazioni e alcuni calcoli è possibile passare tra le due definizioni senza perdere di generalità. La definizione di entropia in questo contesto, così come il metodo di imposizione dei vincoli, verrà approfondita nella sezione 2.5.

2.4 Network Entropy nell'approccio microcanonico

Def. l' **Entropia** di un network è *il logaritmo del numero di network aventi caratteristiche strutturali date*[2] ossia appartenenti all'ensemble generato a partire dal network reale

G. Bianconi definisce esplicitamente il rapporto tra questa misura di entropia e la misura di informazione, asserendo che *minore è l'entropia di un ensemble, maggiore è l'informazione contenuta nei vincoli che lo definiscono.*

È chiaro che aggiungendo vincoli diminuirà il numero di network in grado di soddisfarli, e si prende questo numero come indicatore di specificità dell'ensemble in esame, quindi di contenuto informativo nei vincoli necessari a definirlo.

In questo contesto, un network reale può quindi essere modellizzato per approssimazioni successive imponendo vincoli a partire dal più generale al più specifico. Con la differenza di entropia tra due approssimazioni è possibile quantificare quanto sia restrittiva l'imposizione di uno specifico vincolo.

2.4.1 Imposizione dei vincoli

Ogni vincolo può essere espresso come funzione della matrice di adiacenza. I vincoli possono venire posti, come già accennato, a diversi livelli di profondità, aggiungendo informazione e aumentando la similarità dei network nell'ensemble rispetto l'originale, diminuendone quindi il numero e di conseguenza l'entropia.

La prima approssimazione prevede un ensemble in cui tutti i network hanno lo stesso numero di nodi N e di link L del nostro network di partenza, di fatto un *random graph.*

Una seconda approssimazione può essere fatta imponendo un vincolo sul vettore delle connettività $\vec{k}$, nel qual caso l'ensemble risultante si chiama *configuration ensemble*. Altri vincoli possono essere imposti dando la connettività e la media delle connettività dei primi vicini, o costruendo sotto-comunità di nodi con proprietà particolari e imponendone le relazioni in termini di connessioni.

Noi ci concentreremo anche sul cosiddetto *spatial ensemble*, dove i nodi occupano una posizione in uno spazio geometrico, e viene utilizzata nell'imposizione del vincolo la matrice delle distanze reciproche.

2.4.2 Random graph

L'approssimazione 0 è l'insieme di network con stesso numero di nodi e stesso numero di link:

$$\sum_{i<j} a_{ij} = L \tag{2.3}$$

dove L è la somma di a_{ij} calcolata sul network reale.

In questo caso il numero totale W di network di N nodi che soddisfa il vincolo è facilmente calcolabile come il modo di disporre L nodi in C_2^N link possibili.

$$W_0 = \binom{\frac{1}{2}N(N-1)}{L} \tag{2.4}$$

Si ottiene quindi una misura di entropia al livello 0 di approssimazione attraverso la seguente

$$S_0 = \log(W) = \log \binom{\frac{1}{2}N(N-1)}{L} \tag{2.5}$$

2.4.3 Configuration ensemble

L'approssimazione successiva è l'imposizione di un vincolo sul vettore delle connettività. Per ogni nodo i deve valere

$$\sum_{j} a_{ij} = k_i \tag{2.6}$$

uncorrelated graph Nell'approssimazione di network non-correlato il numero totale di network che soddisfano questa proprietà è ottenibile da considerazioni statistiche.

L'approssimazione che facciamo è la seguente:

$$k_i < \sqrt{\langle k \rangle N} \tag{2.7}$$

Per trovare il numero di network nell'insieme in questione procediamo come segue: assegniamo ad ogni nodo i un numero di *mezze* connessioni pari a k_i, in modo che il problema si riduca nel contare i modi di unire le mezze connessioni tra i nodi. La prima mezza connessione avrà una scelta tra $(2L-1)$ altri mezzi link, e una terza avrà scelta

tra $(2L-3)$ altre mezze connessioni e così via, fino a $(2L-1)!!$ modi di diversi di disporle. Dividendo per l'invarianza che si ottiene permutando internamente le mezze connessioni su ogni nodo, ossia $k_1!k_2!...k_N! = \prod_i k_i!$ si ottiene il numero di modi possibili in cui possiamo disporre le connettività:

$$W'_k = \frac{(2L-1)!!}{\prod_i k_i!} \tag{2.8}$$

Questo approccio tiene però conto della totalità dei grafi che soddisfano i vincoli, inclusi network con multipli link tra gli stessi nodi. Per trattare la presenza di questi network indesiderati, assumiamo che la distribuzione dei match tra le mezze connessioni sia random, ossia che la probabilità di connessione tra due nodi sia governata da una poissoniana, la cui media sia $\frac{k_i k_j}{N\langle k\rangle}$. La probabilità Π, quindi la frazione del totale, che il network non contenga doppi link è quindi stimabile attraverso la distribuzione poissoniana:

$$\Pi = \prod_{i<j} \left(1 + \frac{k_i k_j}{N\langle k\rangle}\right) e^{-\frac{k_i k_j}{N\langle k\rangle}} \simeq e^{-\frac{1}{4}\left(\frac{\langle k^2\rangle}{\langle k\rangle}\right)^2} \tag{2.9}$$

Aggiungendo questo termine alla 2.8 otteniamo il numero di network che soddisfano i vincoli dati sulla connettività, escludendo quelli che presentano due o più connessioni tra gli stessi nodi. L'entropia S è così calcolabile dalla definizione:

$$S_k = -\log W_k = -\log\left(\frac{(2L-1)!!}{\prod_i k_i!} e^{-\frac{1}{4}\left(\frac{\langle k^2\rangle}{\langle k\rangle}\right)^2}\right) \tag{2.10}$$

2.4.4 Spatial ensemble

Oltre al vettore delle connettività, in questo ensemble vengono imposti anche vincoli sul numero di link che stanno in certi intervalli di distanza rispetto una struttura geometrica sottostante.

In questo modello i nodi del network di partenza hanno infatti una coordinata assegnata, $\{\vec{r}_1...\vec{r}_N\}$, dalla quale si estrae informazione sulle distanze nodo-nodo. Questa coordinata può rappresentare una qualunque proprietà dei nodi in questione, variabile ovviamente con la natura fisica del sistema studiato. Nell'esempio dei network di protein-protein interaction (PPI), dove i nodi rappresentano delle proteine di un organismo biologico in vivo e i link rappresentano una connessione funzionale tra le proteine in questione, come ad esempio una collaborazione in una funzione specifica della cellula, o una trasduzione intramembranica di un segnale, la posizione spaziale può venire assegnata a seconda del grado di espressione che le singole proteine esibiscono in un determinato contesto [6], o può venire creata direttamente la matrice delle distanze reciproche misurando la correlazione che i nodi mostrano nei samples di espressione [9].

In ogni caso, l'imposizione del vincolo passa attraverso la creazione della matrice delle distanze reciproche $N \times N$, ovviamente con diagonale nulla. Da questa matrice di distanze relative, il vincolo viene estratto nel seguente modo:

Si decide un numero λ di bin, si distribuiscono le distanze reciproche euclidee $\delta_{ij} = |\vec{r}_i - \vec{r}_j|$ dei nodi tra i bin, e si impone come vincolo il vettore $\vec{B} = \{B_1 ... B_\lambda\}$ del numero di distanze per ogni bin. Attraverso la matrice di adiacenza il vincolo può essere scritto nella forma:

$$\sum_{i<j} a_{ij}\, \chi_l(\delta_{ij}) = B_l \tag{2.11}$$

Dove $\chi_l(\delta)$ è la funzione caratteristica dell'intervallo $l-$esimo, ed è uguale a 1 se $\delta \in B_l$ e 0 altrimenti.

La forma dell'entropia è ottenuta attraverso alcuni calcoli utilizzando un approccio basato sulla funzione di ripartizione Z dell'ensemble e l'approssimazione del punto a sella[1]. La forma trovata è la seguente [3]:

$$\begin{aligned} NS_d = &- \sum_i k_i \omega_i - \sum_l B_l\, g_l \\ &+ \sum_{i<j} \log\left(1 + e^{\omega_i + \omega_j + \sum_l \chi_l(\delta_{ij}) g_l}\right) \\ &- \frac{1}{2}\sum_i \log(2\pi\alpha_i) - \frac{1}{2}\sum_l \log(2\pi\alpha_l) \end{aligned} \tag{2.12}$$

Dove le variabili ω_i e g_l sono i moltiplicatori di Lagrange e soddisfano le equazioni del punto a sella:

$$\begin{aligned} k_i &= \sum_{j \neq i} \frac{e^{\omega_i + \omega_j + \sum_l \chi_l(\delta_{ij}) g_l}}{1 + e^{\omega_i + \omega_j + \sum_l \chi_l(\delta_{ij}) g_l}} \\ B_l &= \sum_{i<j} \chi_l(\delta_{ij})\, \frac{e^{\omega_i + \omega_j + g_l}}{1 + e^{\omega_i + \omega_j + g_l}} \end{aligned} \tag{2.13}$$

Mentre le variabili α_i e α_l sono approssimate dalle espressioni:

$$\alpha_i \simeq \sum_j \frac{e^{\omega_i + \omega_j + \sum_l \chi_l(\delta_{ij}) g_l}}{\left(1 + e^{\omega_i + \omega_j + \sum_l \chi_l(\delta_{ij}) g_l}\right)^2} \tag{2.14}$$

$$\alpha_l \simeq \sum_{i<j} \chi_l(\delta_{ij})\, \frac{e^{\omega_i + \omega_j + g_l}}{\left(1 + e^{\omega_i + \omega_j + g_l}\right)^2} \tag{2.15}$$

I calcoli relativi alle formule qui trovate non verranno approfonditi, in quanto richiedono un formalismo che eccede dai propositi della presente tesi. Tuttavia è interessante

[1] Il formalismo della funzione di partizione canonica Z utilizzato per ricavare questi risultati non è discusso nel presente lavoro. Un approfondimento dettagliato e i calcoli espliciti posso essere trovati in [1, 2, 3]

ottenere un confronto tra i risultati dell'approccio microcanonico, appena discusso, e l'approccio canonico, che approfondiremo nella seguente sezione. Al contrario degli equivalenti in termodinamica, infatti, non è sempre vero che entrambi i metodi convergono ad una soluzione uguale nel limite termodinamico, rappresentato in questo caso da network con un grande numero di nodi e una bassa densità di link.

2.5 Network Entropy nell'approccio canonico

Nell'approccio canonico, come visto nella sezione 2.3.1, si parte dal presupposto che l'ensemble di appartenenza del network reale, che andremo quindi a studiare ad un livello generale, sia descritto da una funzione di probabilità, la cui specifica forma verrà trovata massimizzando l'entropia del network che verrà definita a breve.

La descrizione dell'ensemble qui utilizzata è quella delle probabilità, all'interno dell'ensemble, che due nodi specifici siano siano collegati: p_{ij}. Queste probabilità saranno visualizzate sottoforma di matrice $N \times N$, nella cosiddetta matrice delle probabilità P. In questo contesto possiamo definire l'entropia come

Def. L'**Entropia del Network** nell'approccio canonico è definita, data la matrice di probabilità p_{ij} che descrive l'ensemble di appartenenza del network, come

$$S = -\sum_{i<j} p_{ij} \log(p_{ij}) + (1 - p_{ij}) \log(1 - p_{ij}) \tag{2.16}$$

interpretazione fisica

Da un punto di vista *fisico*, l'entropia in questo approccio, nel caso di network semplici e non diretti, ha la stessa forma dell'entropia di Gibbs di un sistema formato da $N(N-1)$ sottosistemi che possono assumere due stati, lo stato 1 con $P(1) = p_{ij}$ e lo stato 0 con $P(0) = (1-p_{ij})$. L'entropia di Gibbs del sistema diventa quindi la somma delle entropie dei singoli sotto-sistemi, come $\sum_n \sum_s P(s) \log P(s)$, dove n indicizza i sotto-sistemi e s gli stati accessibili a ciascun sottosistema.

interpretazione informazionale

Da un punto di vista della teoria dell'informazione, l'entropia sopra definita è l'entropia di Shannon di una sorgente a $2 \times N(N-1)$ simboli, ciascuno riferito a un link e un non-link tra due nodi specifici, governati dalle probabilità p_{ij} e $(1 - p_{ij})$. Secondo l'interpretazione della teoria dell'informazione, massimizzare l'entropia corrisponde a massimizzare l'ignoranza che abbiamo sul sistema, dichiarando, di fatto, di non fare assunzioni superflue sull'ensemble in questione, considerando i vincoli imposti come unica informazione effettivamente conosciuta all'osservatore. Massimizzare l'entropia significa quindi assumere che il sistema contenga effettivamente *solo l'informazione dovuta ai vincoli*, lasciando ad una distribuzione randomica i restanti gradi di libertà. Questa assunzione è coerente nel momento in cui vogliamo misurare, come effettivamente

viene richiesto, le variazioni di quantità di informazione dovuta all'imposizione specifici vincoli.

2.5.1 Random Network

Nel caso del random network, ossia nell'ensemble di network di N nodi dove solo il numero di link L è imposto come vincolo, è possibile ottenere mediante semplici calcoli statistici la forma della matrice delle probabilità p_{ij}. Siamo infatti nel caso in cui i singoli link nodo-nodo presentano una simmetria rispetto le probabilità, che quindi si riducono ad essere una sola variabile. La forma delle p_{ij} è ricavabile direttamente dal vincolo

$$\phi = \sum_{i<j} p_{ij} - L = 0 \tag{2.17}$$

La massimizzazione mostra esplicitamente questa simmetria. Data la funzione da massimizzare, secondo il formalismo dei moltiplicatori di Lagrange, $\Phi = S + \theta\phi$, si trova infatti

$$\begin{aligned} \frac{\partial \Phi}{\partial p_{ij}} &= \frac{\partial}{\partial\, p_{ij}} \left[-\sum_{i<j} \left(p_{ij}\, \log p_{ij} + (1-p_{ij}) \log(1-p_{ij})\right) + \theta \sum_{i<j} p_{ij} - L \right] \\ &= -\log p_{ij} + \log(1-p_{ij}) + \theta = 0 \\ &\implies p_{ij} = \frac{1}{1+e^{-}\theta} \end{aligned} \tag{2.18}$$

Che mostra come la probabilità sia indipendente dagli indici. A questo punto si potrebbe esplicitare θ con l'equazione del vincolo, ma risulta superfluo in quanto la probabilità $p_{ij} = p$ è facilmente ricavabile:

$$\sum_{i<j} p_{ij} = \sum_{i<j} p = L \implies p = \frac{2L}{N(N-1)} \tag{2.19}$$

L'entropia è quindi:

$$\begin{aligned} S_{rand} &= -\sum_{i<j} p\, \log p \,-\, (1-p) \log(1-p) \\ &= -L \log\left(\frac{2L}{N(N-1)}\right) - \left(\frac{N(N-1)-2L}{2}\right) \log\left(1 - \frac{2L}{N(N-1)}\right) \end{aligned} \tag{2.20}$$

2.5.2 Configuration Ensemble

Come già visto nel caso microcanonico (sezione 2.4.3), il configuration ensemble è l'ensemble definito utilizzando come vincoli il vettore di connettività k_i. l'equazione del vincolo per ogni singolo nodo è:

$$-\sum_{j} p_{ij} + k_i = 0 \quad , \quad i = \{1, \cdots, N\} \tag{2.21}$$

Che in forma di vincolo globale da utilizzare nella massimizzazione dell'entropia è scritta come:

$$\sum_i \phi_i = \sum_i \left(-\sum_j p_{ij} \; + \; k_i \right) \tag{2.22}$$

Il funzionale da massimizzare è quindi:

$$\Phi = S + \sum_i \theta_i \phi_i \tag{2.23}$$

Possiamo riscrivere le equazioni del vincolo come:

$$\begin{aligned} \sum_i \theta_i \phi_i &= \sum_i \theta_i \left(k_i - \sum_j p_{ij} \right) \\ &= \sum_i \theta_i k_i + \sum_{ij} -\theta_i p_{ij} \\ &= \sum_i \theta_i k_i - \sum_{i<j} (\theta_i + \theta_j) p_{ij} \end{aligned} \tag{2.24}$$

L'espressione delle p_{ij} che massimizzano l'entropia è quindi calcolabile:

$$\begin{aligned} \frac{\partial \, \Phi}{\partial p_{ij}} &= -\log p_{ij} + \log(1 - p_{ij}) - (\theta_i + \theta_j) = 0 \\ \implies p_{ij} &= \frac{1}{1 + e^{(\theta_i + \theta_j)}} = \frac{e^{-(\theta_i + \theta_j)}}{1 + e^{-(\theta_i + \theta_j)}} \end{aligned} \tag{2.25}$$

Le variabili θ_i devono soddisfare le equazioni dei vincoli, e si possono trovare risolvendo il sistema a N equazioni

$$\sum_j \frac{1}{1 + e^{(\theta_i + \theta_j)}} = k_i \quad , \quad i = \{1, \cdots, N\} \tag{2.26}$$

Chiaramente non sempre è possibile affrontare analiticamente il sistema, che viene quindi risolto computazionalmente, trovando i valori specifici dei moltiplicatori con una precisione a piacere.

2.5.3 Spatial Ensemble

Lo spatial ensemble è definito ponendo come vincoli il vettore di connettività $\vec{k}$ e il numero di link nodo-nodo che collegano due nodi in certi intervalli di distanza. La definizione dei vincoli è la stessa dell'approccio microcanonico (sezione 2.4.4): si decide un numero λ di bin in cui dividere la minima e la massima distanza tra due nodi esibita dal network, si contano quante distanze euclidee $\delta_{ij} = |\vec{r_i} - \vec{r_j}|$ relative a due nodi linkati appartengono ad ogni specifico bin di distanza e si impongono λ vincoli riferiti alla distribuzione risultante sui bin $\vec{B} = \{B_1 ... B_\lambda\}$. Formalmente abbiamo quindi le equazioni che descrivono i vincoli come

$$\phi_i = \sum_j p_{ij} = k_i \quad i = \{1, \cdots, N\} \tag{2.27}$$
$$\phi_l = \sum_{i<j} p_{ij}\, \chi_l(\delta_{ij}) - B_l = 0 \quad l = \{N+1, \cdots, N+\lambda\}$$

Dove $\chi_l(\delta)$ è la funzione caratteristica dell'intervallo $l-$esimo, ed è uguale a 1 se $\delta \in B_l$ e 0 altrimenti. In totale avremo quindi N vincoli dovuti al vettore connettività e λ vincoli dovuti alla distribuzione spaziale. Il funzionale da massimizzare sarà quindi

$$\Phi = S + \sum_i \theta_i\, \phi_i + \sum_l \theta_l\, \phi_l$$

Esplicitamente:

$$\begin{aligned} \Phi = &- \sum_{i<j} \left(p_{ij}\, \log p_{ij} + (1-p_{ij}) \log(1-p_{ij})\right) \\ &+ \sum_{i=1}^{N} \theta_i \left(k_i - \sum_j p_{ij}\right) \\ &+ \sum_{l=N+1}^{N+\lambda} \sum_{i<j} \theta_l \left(B_l - \chi_l(\delta_{ij}) p_{ij}\right) \end{aligned} \tag{2.28}$$

Attraverso la massimizzazione l'entropia, differenziando con il formalismo di Lagrange, troviamo quindi un'espressione per le probabilità p_{ij} :

$$\begin{aligned} 0 &= \frac{\partial}{\partial\, p_{ij}} \left(S + \sum_i \theta_i\, \phi_i + \sum_l \theta_l\, \phi_l\right) \\ &= -\log p_{ij} + \log(1-p_{ij}) - (\theta_i + \theta_j) - \sum_l \chi_l(\delta_{ij})\theta_l \\ \implies p_{ij} &= \frac{e^{-\left(\theta_i + \theta_j + \sum_l \chi_l(\delta_{ij})\theta_l\right)}}{1 + e^{-\left(\theta_i + \theta_j + \sum_l \chi_l(\delta_{ij})\theta_l\right)}} \end{aligned} \tag{2.29}$$

Possiamo riscrivere questo risultato come:

$$p_{ij} = \sum_l \chi_l(\delta_{ij}) \frac{e^{-(\theta_i+\theta_j+\theta_l)}}{1 + e^{-(\theta_i+\theta_j+\theta_l)}} \tag{2.30}$$

2.6 Single Node Entropy S_i

2.6.1 Definizione

L'entropia di singolo nodo è definita come l'entropia di Shannon del vettore corrispondente al nodo nella matrice delle probabilità p_{ij} caratteristica dell'ensemble.

Per considerare il detto vettore i-esimo come una sorgente caratterizzabile mediante la misura di Shannon, è necessario normalizzare le probabilità in modo da avere $\sum_{j=1}^{N} p'_{ij} = 1$. Dividendo quindi per k_i otteniamo la definizione:

Def. La **Single Node Entropy** S_i è definita come:

$$S_i = -\sum_{j \neq i} p'_{ij} \log p'_{ij} \tag{2.31}$$

Dove $p'_{ij} = \frac{p_{ij}}{k_i}$ sono le probabilità di link normalizzate sulla riga relativa al nodo i-esimo.

2.6.2 Definizione alternativa S_i^{TS}

In letteratura è definita un'altra misura di entropia per il singolo nodo, utilizzata da Teschendorff e Severini [9] come misura dei cambiamenti nel flusso dell'informazione che avvengono in una neoplasia metastatica rispetto alla situazione non metastatica.

In questa definizione manca il contesto di ensemble visto precedentemente, e la costruzione della matrice p_{ij}, che verrà poi utilizzata nell'entropia, si fonda sull'assunto che due nodi che mostrano una correlazione maggiore a livello di espressione hanno più probabilità di essere collegati, ossia, nell'interpretazione dell'articolo, di essere proteine mutualmente interagenti al fine di una funzione cellulare specifica.

Al contrario di quella presentata in questo lavoro, questa misura non può prescindere dall'esistenza di una metrica tra i nodi e quindi di una matrice $N \times N$, C_{ij}, la quale assume, in un certo senso, lo stesso ruolo della matrice di distanza nell'ambito degli spatial ensemble.

Per trovare quindi matrice delle p_{ij} si procede nel seguente modo: si assegna ai nodi connessi, secondo la matrice delle adiacenze, un peso sulla connessione $w_{ij} = C_{ij}$ dove C_{ij} è l'indice di correlazione di Pearson tra i due samples relativi ai nodi. Essendo $C_{ij} \in [-1, 1]$ si trasformano gli indici linearmente in modo da avere valori in $[0, 1]$, ossia $w_{ij} = \frac{1}{2}(C_{ij} + 1)$.

A questo punto si normalizzano i valori in modo da avere la somma unitaria sulla riga, trovando quindi la matrice p_{ij}, ossia:

$$p_{ij} = \frac{w_{ij}}{\sum_{j \in N(i)} w_{ij}} \tag{2.32}$$

Dove $N(i)$ rappresenta l'insieme degli indici j tali che $a_{ij} = 1$.

L'entropia di singolo nodo è quindi definita come l'entropia di Shannon della riga della matrice p_{ij} relativa al nodo, normalizzata secondo la connettività:

Def. L'entropia di singolo nodo S^{TS} è definita come:

$$S_i^{TS} = -\frac{1}{\log(k_i)} \sum_{j \in N(i)} p_{ij} \log p_{ij} \tag{2.33}$$

Sebbene abbiano una forma molto simile, i comportamenti della S_i e della S^{TS} presentano differenze sostanziali, dovute principalmente al diverso contesto in cui sono state definite. Le analogie e le differenze tra queste due misure saranno gli argomenti principali del capitolo 3, in cui verranno messe a confronto per evidenziarne le caratteristiche oltre che la capacità di fornire informazione sull'appartenenza dei singoli nodi a situazioni differenti nell'ambito degli spatial ensemble.

Capitolo 3

Caratterizzazione delle misure di Single Node Entropy

3.1 S_i nel Configuration Ensemble

In questa sezione si mostrano alcune proprietà della misura di Single Node Entropy definita nel presente lavoro, esibite nel contesto dei configuration ensemble, ossia utilizzato il vettore di connettività come vincolo nella definizione della matrice dell'ensemble. Nel caso delle entropie configurazionali è possibile effettuare alcune considerazioni analitiche sulla misura, in ragione anche di alcune approssimazioni che semplificano notevolmente i calcoli.

3.1.1 distribuzione delle S_i all'interno dell'ensemble

Una prima analisi può essere effettuata sulle entropie di singolo nodo a livello statistico, ossia nel merito della forma della loro distribuzione nell'ensemble di appartenenza.

Nel limite di *uncorrelated graph* è possibile dare un'espressione analitica del valore di S_i in funzione della connettività. Abbiamo visto infatti, nella sezione 2.4.3, che nel caso configurazionale dei network, quando vale la condizione $k_i < \sqrt{N \langle k \rangle}$, si può approssimare la singola p_{ij} come

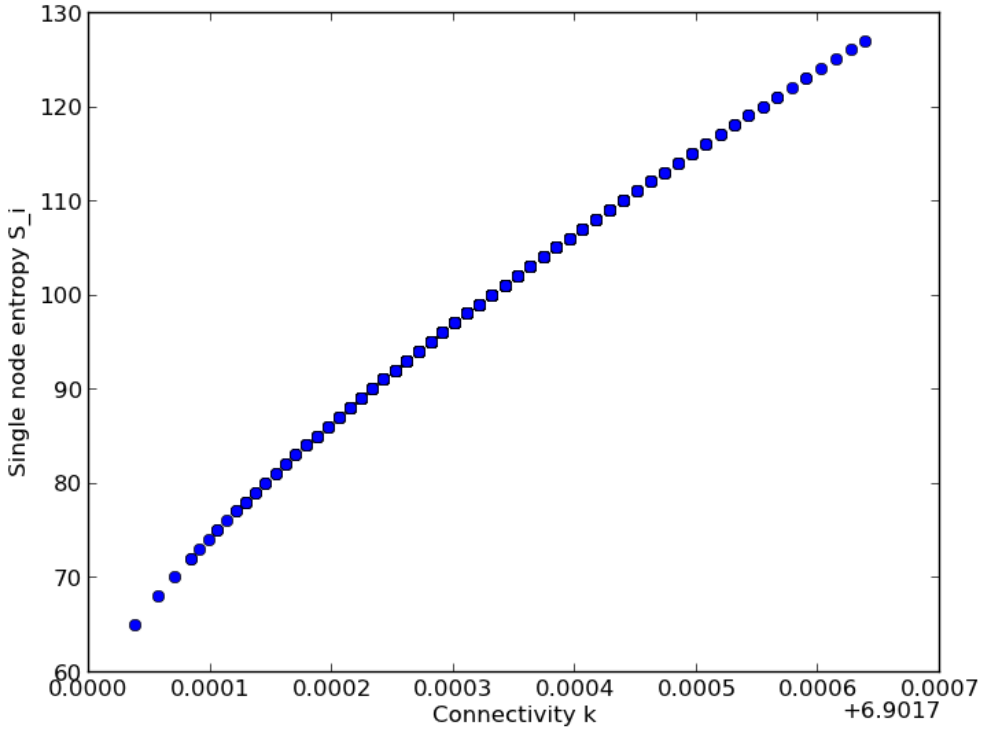
$$p_{ij}^{unc} = \frac{k_i k_j}{N \langle k \rangle} \tag{3.1}$$

L'accuratezza di questa approssimazione nel presente contesto è stata verificata confrontando i valori di entropia così approssimati con quelli effettivamente calcolati. Si è osservato che, con un network di 10000 nodi, l'errore relativo massimo tra i valori calcolati e i valori reali è dell'ordine di $\simeq \frac{0.001}{9.2} \simeq 1 \cdot 10^{-4}$. Precisamente si è trovato infatti $max(\frac{S_i - S_i^{app}}{S_i}) = 1,0012 \cdot 10^{-4}$

L'andamento dell'entropia in funzione della connettività è mostrato nella figura 3.1

Verificata la coerenza dell'approssimazione possiamo sostituire 3.1 nella definizione di S_i. otteniamo quindi:

Figura 3.1: Plot di S_i in funzione di k; $N = 1000$, $p_{link} = .1$



$$
\begin{aligned}
S_i &= -\sum_{j \neq i} \frac{k_i k_j}{k_i \cdot N \langle k \rangle} \log \frac{k_i k_j}{k_i \cdot N \langle k \rangle} \\
&= -\sum_{j=1}^{N} \frac{k_j}{N \langle k \rangle} \log \frac{k_j}{N \langle k \rangle} + \frac{k_i}{N \langle k \rangle} \log \frac{k_i}{N \langle k \rangle} \\
&= S_0 + \frac{k_i}{N \langle k \rangle} \log \frac{k_i}{N \langle k \rangle}
\end{aligned}
\tag{3.2}
$$

Dove S_0 è una costante indipendente dal nodo scelto e corrisponde all'entropia di Shannon del vettore connettività normalizzato.

Risulta quindi che l'entropia di singolo nodo così definita, nel caso in cui sia valida l'approssimazione 3.1, è una funzione precisa della connettività k_i.

È naturale aspettarsi quindi che la distribuzione delle entropie sarà della stessa forma di quella delle connettività, che nell'approccio canonico sono governate da una distribuzione binomiale.

Le k_i rappresentano infatti, dal punto di vista della costruzione algoritmica della matrice delle adiacenze, il risultato dell'iterazione di una probabilità a due stati, in cui è data la probabilità associata al link $P(a_{ij} = 1) = p_{link}$.

Essendo il numero di nodi abbastanza grande, è coerente approssimare questa distribuzione ad una gaussiana, distribuzione che sarà quindi seguita anche dalle entropie di

singolo nodo in questione. La forma della distribuzione delle connettività e delle entropie è mostrata in figura 3.2. In questo plot si è utilizzata una matrice con $N = 10000$, e la probabilità p_{link} è stata impostata a 0.2. Si è infatti ottenuta, per le connettività, una gaussiana centrata in $N \cdot p_{link} = 2000$, mentre per le S_i la distribuzione è centrata su un valore prossimo a $\log N$.

Figura 3.2: Distribuzione delle entropie di singolo nodo S_i all'interno dell'ensemble configurazionale; $N = 10000$

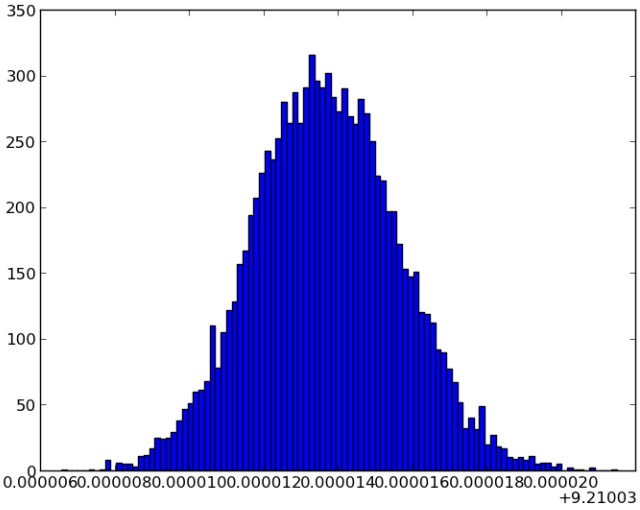
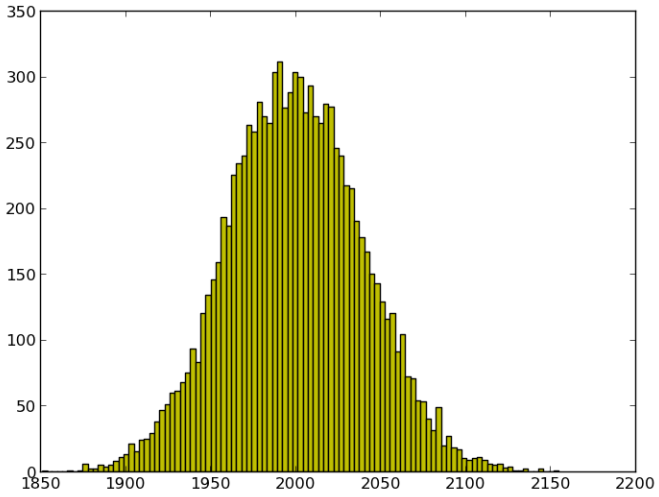


Figura 3.3: Distribuzione delle connettività k_i



3.1.2 Sensibilità alla connettività media $\langle k_i \rangle$

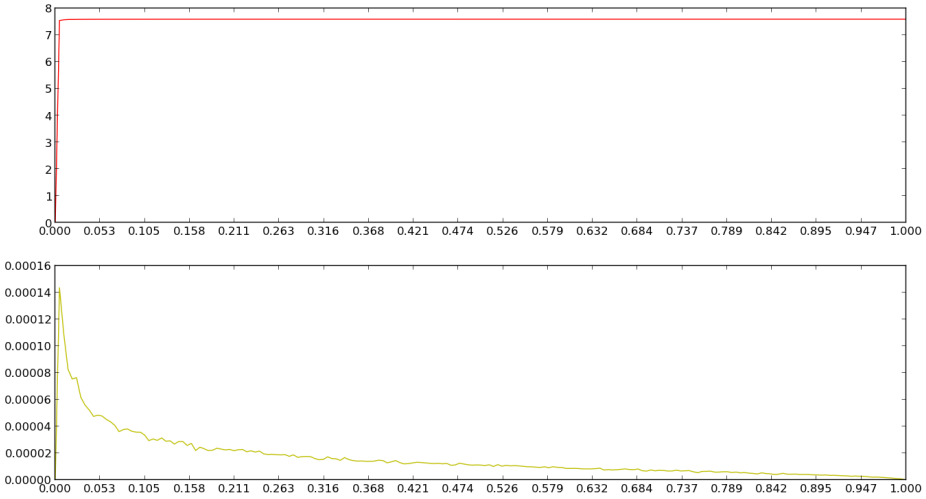
Al contrario dell'entropia S globale del Network, l'entropia di singolo nodo, in termini di media, non rappresenta una funzione simmetrica al variare della p_{link}, probabilità di

generare un link nella generare la matrice delle adiacenze a_{ij} *madre* dell'ensemble.

L'andamento della media di S_i al variare della p_{link} è mostrato in figura 3.4.

Le matrici sono state generate con $N = 2000$. Ogni valore di media e deviazione standard è calcolato su una singola istanza, cioè su 2000 valori. Si nota che i valori sono stabili intorno a $log(2000) \simeq 7,6$.

Figura 3.4: Media delle S_{SN} al variare della p_{link} (in alto) e deviazione standard relativa (in basso)



Possiamo calcolare la media delle entropie di singolo nodo in un ensemble canonico costruito a partire da una matrice generata con una probabilità di link p_{link}. Questa probabilità è strettamente correlata alla connettività media, in quanto definisce la distribuzione delle k_i che verranno successivamente utilizzate come vincolo nella generazione dell'ensemble, come visto nella sezione 3.1.1.

Per la definizione stessa di ensemble, infatti, il valore medio delle connettività deve risultare uguale alla probabilità di singolo link per il numero di nodi: $\langle k_i \rangle = Np$ dove $p = p_{link}$. Mettendoci in queste condizioni, possiamo sostituire nell'equazione 3.2 per trovare, in approssimazione di *uncorrelated graph*, un'espressione analitica della media delle S_i.

Consideriamo innanzitutto che $N \langle k \rangle = \sum_i k_i = N^2 p$, definiamo quindi $k'_i = k_i / \sum_i k_i = k_i / N^2 p$, le connettività normalizzate.

La media delle entropie di singolo nodo può essere espressa come:

$$\begin{aligned}\langle S_i\rangle &= \left\langle -\sum_{j=1}^{N} \frac{k_j}{N\langle k\rangle}\log(\frac{k_j}{N\langle k\rangle}) + \frac{k_i}{N\langle k\rangle}\log(\frac{k_i}{N\langle k\rangle})\right\rangle \\ &= \left\langle -\sum_{j=1}^{N} k'\log k'\right\rangle + \langle k'\log k'\rangle \\ &= N\langle y\rangle + \langle y\rangle\end{aligned} \tag{3.3}$$

Dove y è la variabile distribuita come $y = \frac{k_i}{N\langle k\rangle}\log\frac{k_i}{N\langle k\rangle}$. La media di questa distribuzione può essere trovata utilizzando l'assunto dell'approssimazione a gaussiana della binomiale che governa la distribuzione delle connettività.

$$\langle y\rangle = \int_1^\infty \frac{x}{N^2p}\log\frac{x}{N^2p}\cdot G_{\mu,\sigma^2}(x)dx \tag{3.4}$$

Nell'approssimazione del limite binomiale $\rightarrow$ gaussiana, la distribuzione normale risultante ha $\mu = Np$ e $\sigma = \sqrt{Np(1-p)}$. Possiamo riscalare per N^2p la distribuzione delle connettività in modo da trovare la distribuzione di k':

$$G(Np,\, Np(1-p)) \rightarrow G(\frac{1}{N}, \frac{(1-p)}{N^3p}) \tag{3.5}$$

Otteniamo quindi:

$$\langle y\rangle = \frac{\sqrt{N^3p}}{\sqrt{2\pi\,(1-p)}}\int_1^\infty x\log x\cdot e^{-\frac{N^3p}{1-p}(x-\frac{1}{N})^2}dx \tag{3.6}$$

Purtroppo questo integrale è molto difficile da risolvere per via analitica, ma è possibile, se necessario, calcolarlo conputazionalmente con un'approssimazione a piacere.

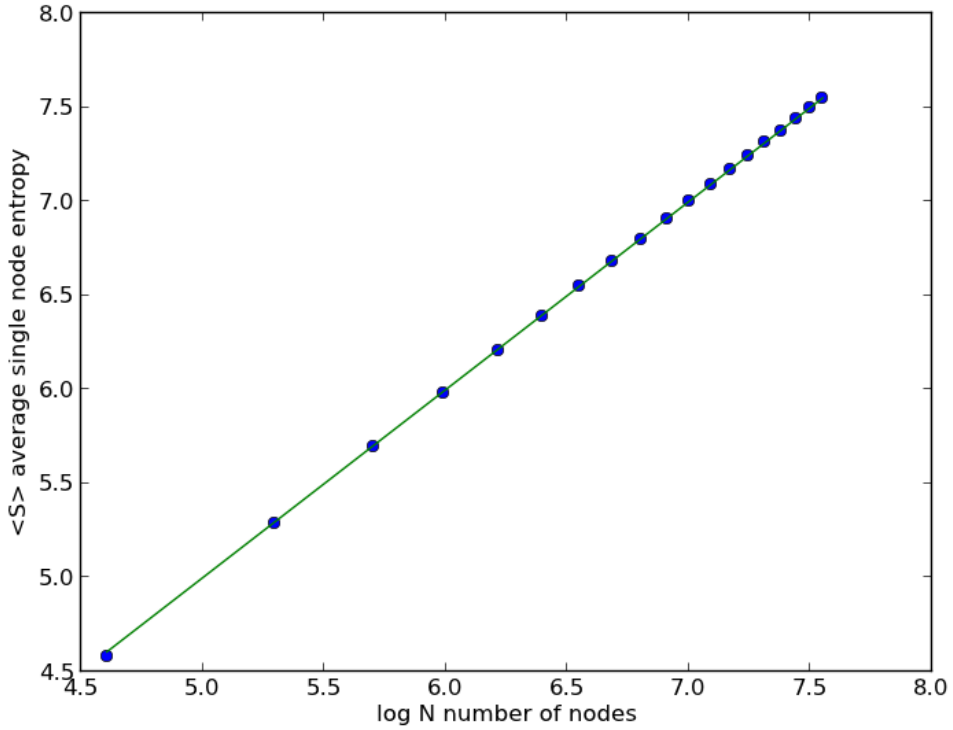
campo medio Nell'approssimazione di *campo medio*, tuttavia, è possibile ottenere un'espressione analitica della media delle entropie in funzione del numero di nodi N. In questa approssimazione sostituiamo $\langle f(x)\rangle\, conf(\langle x\rangle)$ nella 3.3 e otteniamo:

$$\begin{aligned}\langle S_i\rangle &= \left\langle -\sum_{j=1}^{N} \frac{k_j}{N\langle k\rangle}\log(\frac{k_j}{N\langle k\rangle}) + \frac{k_i}{N\langle k\rangle}\log(\frac{k_i}{N\langle k\rangle})\right\rangle \\ &\simeq -\sum_{j=1}^{N} \frac{\langle k_j\rangle}{N\langle k\rangle}\log(\frac{\langle k_j\rangle}{N\langle k\rangle}) + \frac{\langle k_j\rangle}{N\langle k\rangle}\log(\frac{\langle k_j\rangle}{N\langle k\rangle}) \\ &= -\sum_{j=1}^{N} \frac{1}{N}\log\left(\frac{1}{N}\right) + \frac{1}{N}\log\left(\frac{1}{N}\right) \\ &= (1-\frac{1}{N})\log N\end{aligned} \tag{3.7}$$

Questo risultato spiega quindi l'indipendenza dalla connettività media della media dei valori di entropia di singolo nodo, mostrata dal modello in figura 3.4. Mostra inoltre

che questo valore medio si approssima a $\log N$ per valori grandi di N, valore verificato in figura 3.5

Figura 3.5: Valori di media delle entropie di singolo nodo in funzione del logaritmo di N, con il relativo fit lineare.



3.2 Spatial Ensemble

La caratterizzazione e il confronto delle misure S_i e S_i^{TS} sono state affrontate nell'ambito dei network spaziali nel formalismo canonico (si veda sezione 2.5.3). Come già visto, le soluzioni analitiche sono molto difficili da trovare in questo contesto, e i risultati sono ottenuti mediante metodi di approssimazione computazionali.

È stata perciò realizzata una serie di modelli atti a caratterizzare la misura in condizioni diverse, variando la disposizione spaziale e misurando volta per volta le caratteristiche dell'ensemble generato.

Ogni modello è stato realizzato mantenendo costante la matrice delle adiacenze a_{ij}, di 1000 nodi generata con $p_{link} = .1$, e variando in modo diverso la posizione che i nodi vanno ad occupare nella sottostruttura geometrica.

Ogni misura in situazioni spaziali non banali può quindi essere confrontata con una *zero situation*, variabile da modello a modello che consiste in una situazione di *riposo* delle occupazioni spaziali, e con una *zero-zero situation*, ossia l'entropia di singolo nodo in assenza di sottostruttura spaziale (caso configurazionale), mantenendo ovvia-

mente invariata la matrice delle adiacenze di partenza. Le differenze punto-punto tra le entropie spaziali variate e quelle configurazionali possono essere usate per palesare la variazione dovuta ai vincoli ricavati dalla metrica tra i nodi, appiattendo la tipica dipendenza dei valori di S_i dalle specifiche connettività.

3.2.1 Comportamento *entropico*

Come prima analisi è stato studiato il comportamento delle due misure al variare della distribuzione dei valori della matrice di distanze. Nello specifico, è stata misurata la media delle entropie di singolo nodo, in entrambi i casi, al variare per intervalli della randomness che governa i valori nella matrice di distanza.

La situazione di partenza prevede una matrice $N \times N$ di valori identici, presi a 0.5, e al variare di un parametro r una percentuale sempre maggiore di questi valori viene sostituita da valori random compresi tra 0 e 1. In questo modo si è voluto vedere quanto il comportamento delle misure sia correlato all'incertezza nella distribuzione dei valori nella matrice di distanza, caratterizzandone in un certo modo la natura *entropica* in riferimento a questa matrice.

Entropia di Teschendorff e Severini S^{TS}

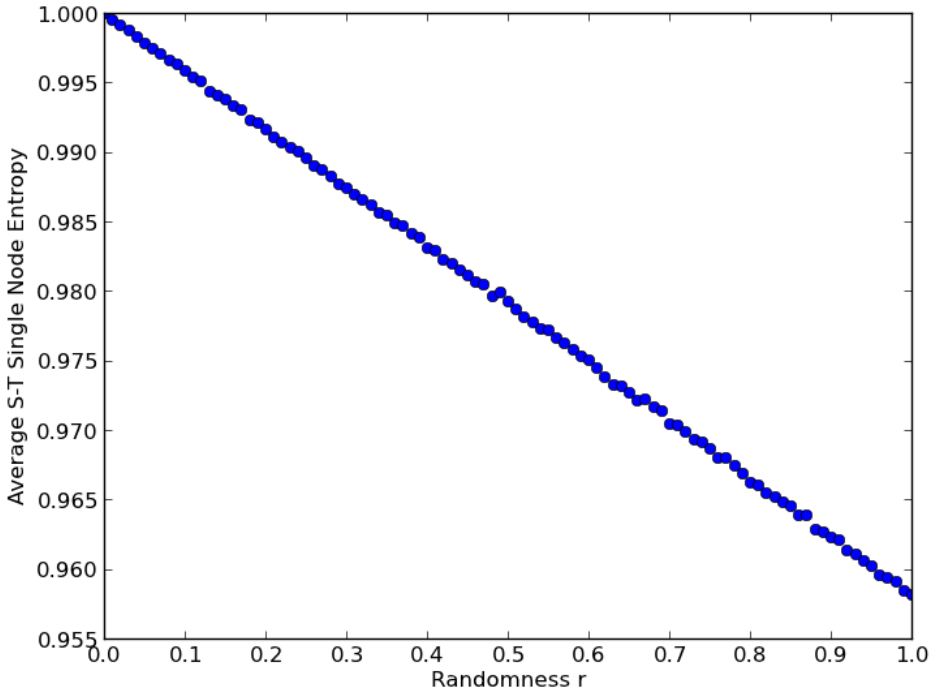
Dalla definizione dell'entropia di singolo nodo di Teschendorff e Severini è evidente che ci si aspetta un comportamento molto simile all'entropia di Shannon rispetto alla matrice delle distanze. Mancando infatti la generazione della matrice delle probabilità attraverso l'approccio canonico, S_i^{TS} viene misurata attraverso l'entropia dell'informazione proprio sui valori della matrice di distanze. Nel caso estremo, $r = 0$, i valori sono tutti uguali, e le p_{ij} saranno di conseguenza valori identici normalizzati sul vettore del nodo tali da avere somma unitaria, ossia $p_{ij} = \frac{1}{k_i}$. Si trova quindi

$$S_i^{TS} = -\frac{1}{\log k_i} \sum_{j \in N(i)} p_{ij} \log p_{ij} = -\frac{1}{\log k_i} \sum_{j \in N(i)} \frac{1}{k_i} \log \frac{1}{k_i} = 1 \tag{3.8}$$

Ossia il valore massimo assumibile dalla misura. Nell'interpretazione informazionale, infatti, questo caso rappresenta la distribuzione più disordinata possibile, in cui abbiamo il quantitativo minimo di informazione: tutti i simboli hanno uguale probabilità, quindi non possiamo fare nessun tipo di considerazione che permetta di gestire le uscite dei simboli in modo ottimale. È il caso analogo, ad esempio, del dado non truccato, in cui ogni faccia ha esattamente la stessa probabilità di essere mostrata dopo un lancio. Il comportamento della misura al variare del parametro r è mostrato in figura 3.6.

Nel secondo caso estremo, $r = 1$, la distribuzione presenta una grande varietà di valori, che corrisponde ad una entropia informazione minore. In questo caso non si osserva un forte calo di entropia, in quanto i valori random sono distribuiti in modo non ottimale dal punto di vista dell'informazione: il minimo, 0, sarebbe raggiunto solo nel caso in cui si avesse un valore di p_{ij} molto prossimo a 1 e tutti gli altri molto prossimi a 0,

Figura 3.6: Comportamento della media delle entropie di singolo nodo S_i^{TS} al variare della randomness nella matrice di distanza



caso in cui l'informazione contenuta nella distribuzione, o ignoranza che l'osservatore ha nei suoi confronti, è praticamente nulla, da cui la misura nulla di entropia di Shannon.

Single Node Entropy S_i

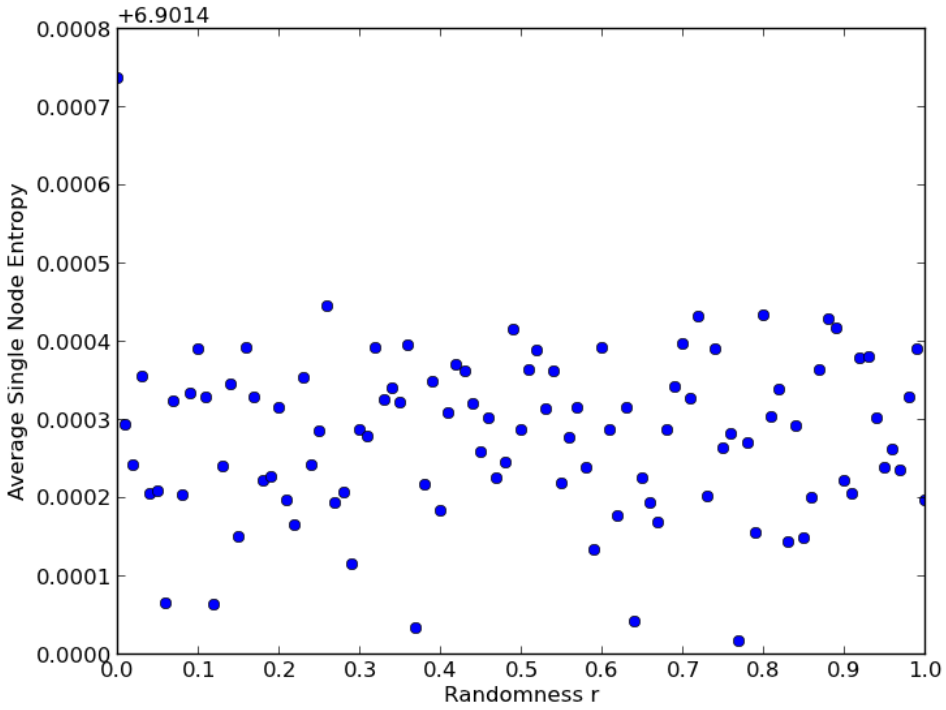
Nell'approccio canonico alla generazione della matrice p_{ij} è molto difficile fare considerazioni analoghe a quelle viste per l'entropia TS. È stato quindi graficato l'andamento della media in funzione di r, figura 3.7, mostrando sostanzialmente un comportamento invariante.

Le p_{ij}, infatti, non sono sottoposte a vincoli in modo diretto a partire dalla matrice di distanze, ma sono trovate massimizzando l'entropia ponendo vincoli globali sul network, ossia raccogliendo i link della matrice in una serie di bin di distanze e ponendo come vincolo canonico la conservazione del numero di link per ogni bin.

3.2.2 Toy models per la Single Node Entropy

Nei modelli qui esposti, la $0-situation$ prevede che ai nodi sia assegnata una coordinata spaziale secondo una distribuzione gaussiana $G_0 = G(\mu_0, \sigma_0)$ con media e deviazione standard assegnati. La distribuzione delle entropie di singolo nodo nella situazione a riposo è mostrata in figura 3.8

Figura 3.7: Comportamento della media delle entropie di singolo nodo S_i al variare della randomness nella matrice di distanza



Ogni modello prevede un confronto tra questa situazione, detta *a riposo*, e una situazione *variata* in cui la prima metà dei nodi viene distribuita secondo G_0, mentre la posizione della seconda metà dei nodi è governata da una distribuzione gaussiana G_{var} con diversa media o diversa deviazione standard.

Le posizioni così distribuite vengono quindi utilizzate per creare una matrice di distanze reciproche, che verrà utilizzata come vincolo nel calcolo di massimizzazione dell'entropia. L'ensemble così trovato è quindi descritto dalla matrice p_{ij}, da cui vengono calcolate le entropie di singolo nodo secondo la definizione 2.31

Ogni modello è costituito da un codice in Python per generare le matrici di distanza, le matrici di adiacenza e per l'elaborazione grafica dei risultati, e utilizza un algoritmo in C++ per il calcolo della matrice dell'ensemble e delle entropie di singolo nodo. In questo modo si è ottenuto un miglioramento dal punto di vista dei tempi di calcolo, senza tuttavia influire sulla praticità di elaborazione ad alto livello caratteristica del Python.

Il confronto è avvenuto separatamente per i modelli in cui variano deviazione standard e quelli in cui varia la media, in modo da poter analizzare indipendentemente come i due parametri governano una eventuale separazione tra le entropie di singolo nodo risultanti.

L'obiettivo del confronto è trovare un osservabile che discrimini i nodi appartenenti al

Figura 3.8: Valori di S_i nella situazione spaziale a riposo con $\mu_0 = 1$ e $\sigma_0 = 1$. In rosso i nodi appartenenti alla prima metà, in giallo quelli nella seconda metà. Non si osserva chiaramente differenziazione tra i due gruppi.

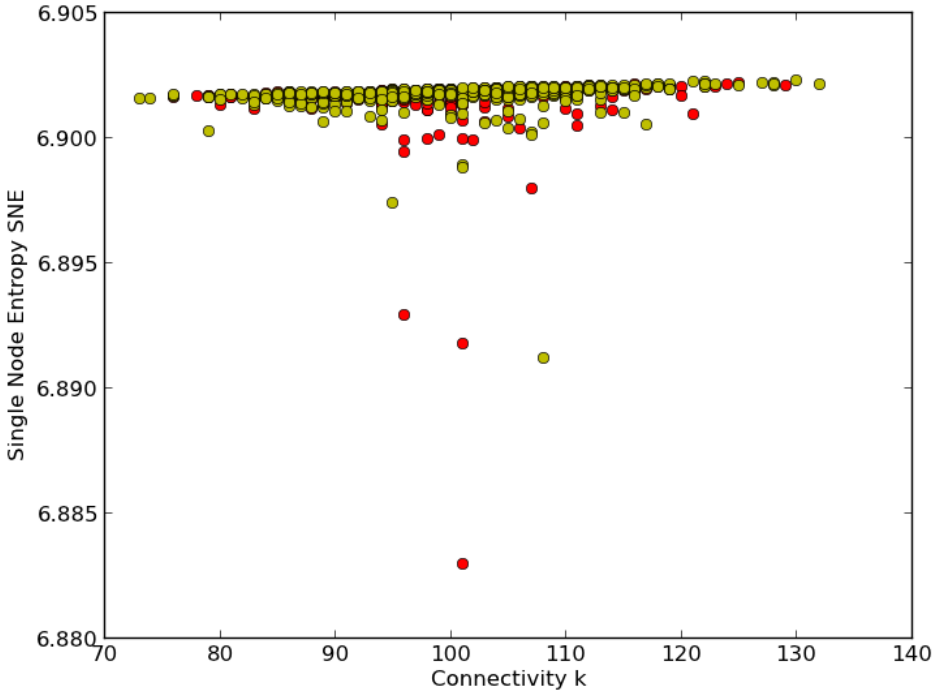
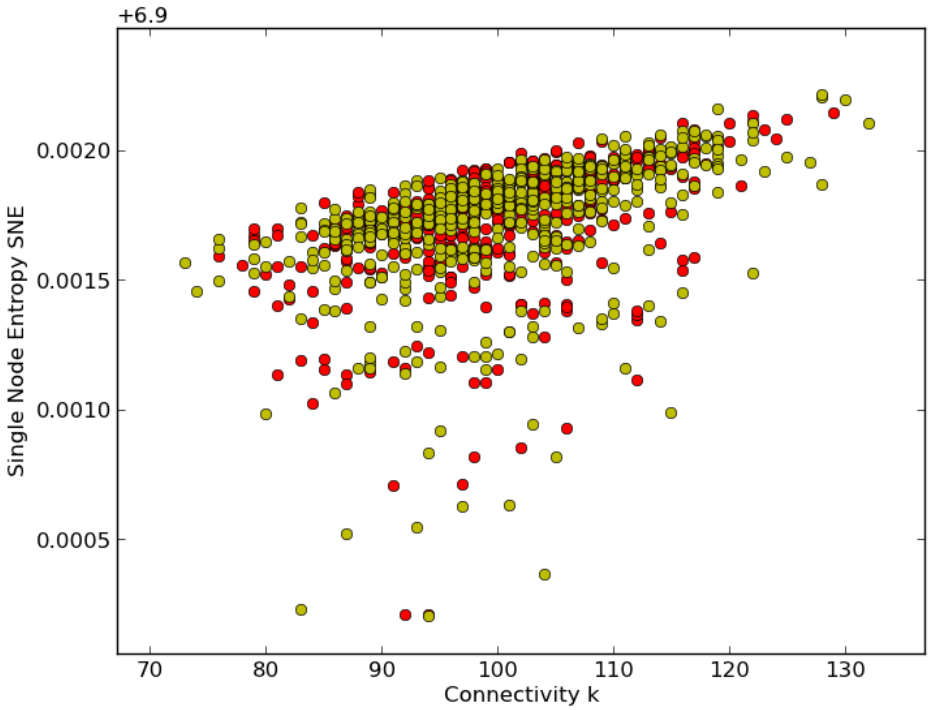


Figura 3.9: Zoom della distribuzione



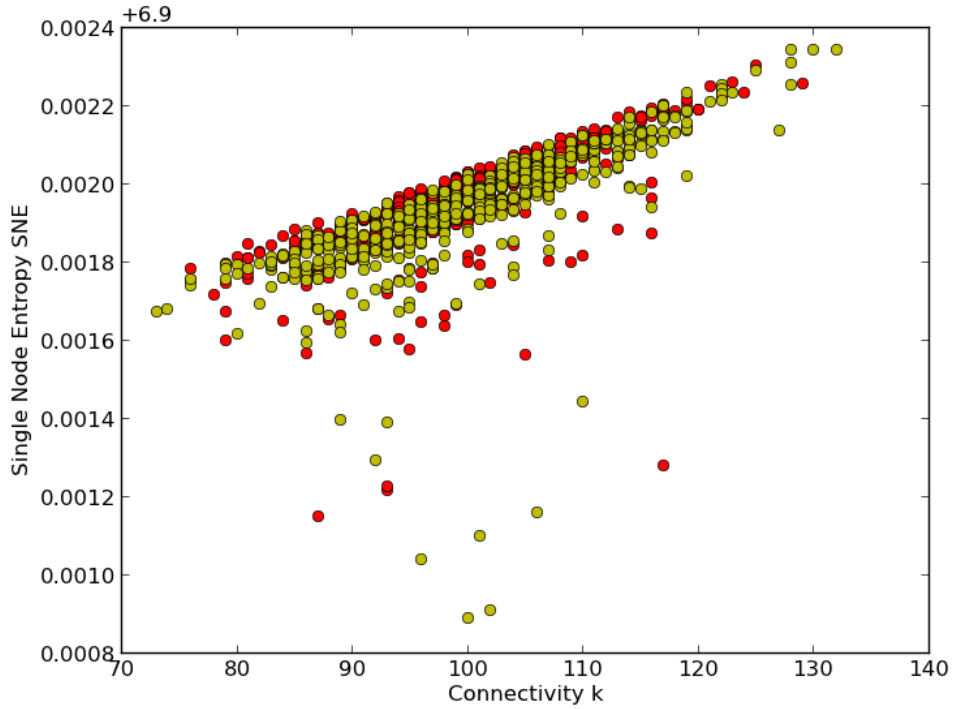
secondo gruppo, G_{var}, da quelli appartenenti al primo. Per far ciò sono state confrontate media e deviazione standard all'interno dei due sotto-gruppi di nodi, ed è stata fatta un'analisi grafica plottando le due distribuzioni in colori separati. Si è quindi analizzata allo stesso modo la distribuzione delle differenze punto-punto ΔS^0 delle entropie tra le due distribuzioni.

3.2.3 μ model

Il primo modello consiste nel variare sensibilmente la media nella distribuzione che governa la posizione spaziale della seconda metà dei nodi, ossia $G_\mu = G(\mu_{var}, \sigma_0)$, mantenendo G_0 come distribuzione della prima metà.

In figura 3.10 sono state plottate le entropie di singolo nodo risultanti da un modello variato con $\sigma = \sigma_0$, $\mu_{var} = 100 \cdot \mu_0$, in funzione della connettività. In rosso i nodi $\in G_0$, in giallo i nodi $\in G_\mu$.

Figura 3.10: Valori di S_i nella situazione spaziale variata μ-**model** con $\mu_{var} = 100 \cdot \mu_0$ e $\sigma_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_μ



In figura 3.11 si sono plottate le differenze punto-punto ΔS^0 delle entropie di singolo nodo tra la situazione spaziale variata in questione e quella a riposo.

considerazioni In termini di separazione non sembra esserci differenza significativa tra le due distribuzioni. Questo risultato è interpretabile alla luce delle distribuzioni

Figura 3.11: Valori di ΔS^0 tra la situazione spaziale variata μ-**model** con $\mu_{var} = 100$ e la situazione a riposo.

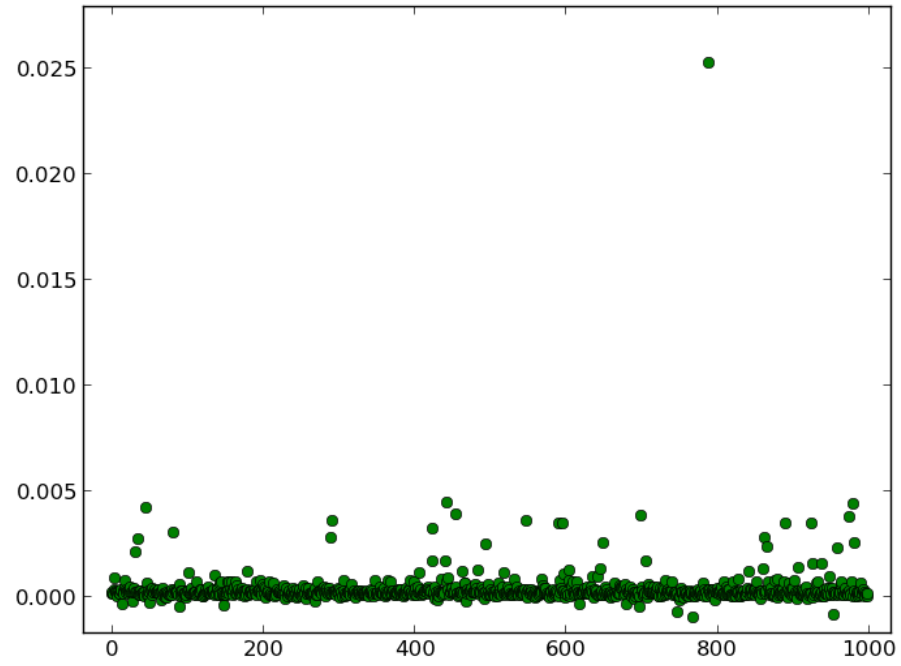


Figura 3.12: Zoom sulla distribuzione

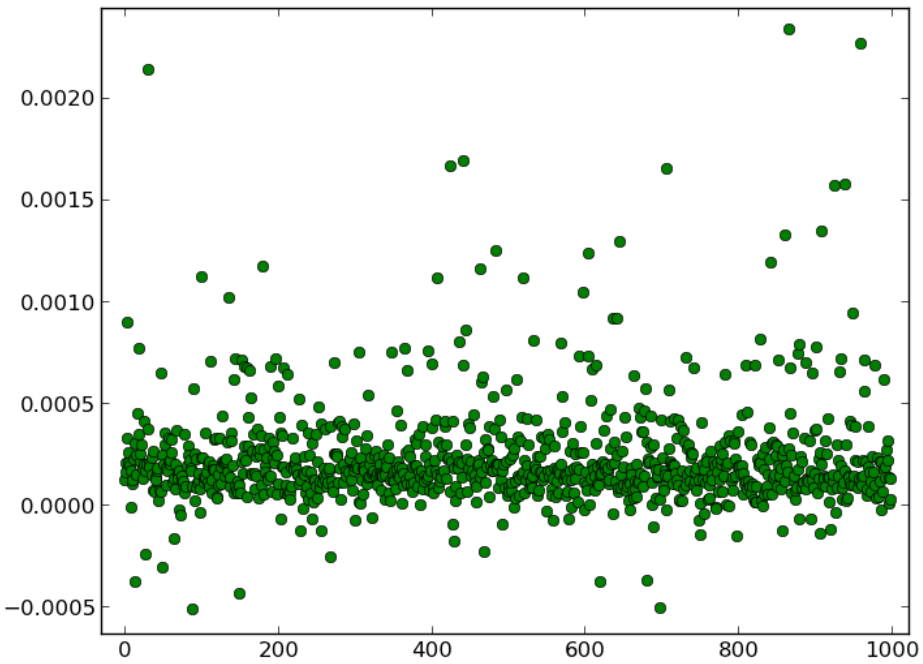


Tabella 3.1: Valori di media deviazione standard per i due gruppi di nodi nel μ**-model**

μ **-model**	media	deviazione standard
globale	6.901945	0.000153
nodi $\in G_0$	6.901955	0.000137
nodi $\in G_\mu$	6.901936	0.000167
globale 0-sit	6.901600	0.001374

spaziali governanti i sottogruppi di nodi: siamo infatti in un caso simmetrico, in cui le due distribuzioni non hanno differenze particolari a meno della distanza reciproca a cui sono poste. È quindi naturale aspettarsi che nelle entropie, calcolate a partire da vincoli sulle distanze relative, non si osservi una differenziazione.

Questa spiegazione trova conferma, come mostrato in tabella 3.1, dai valori di media e deviazione standard delle entropie nei sotto-gruppi di nodi in questione, in cui non si osservano variazioni sensibili.

È tuttavia osservata una diminuzione sensibile della deviazione standard, globale e dei sottogruppi, rispetto alla situazione 0. Una spiegazione del fenomeno può essere data in ragione del metodo di imposizione dei vincoli nel calcolo della massima entropia nel trovare la matrice p_{ij} dell'ensemble. Nel procedimento, infatti, le distanze reciproche vengono raggruppate per bin, con un numero fissato di bin in funzione di N, al fine di ottenere una distribuzione delle distanze reciproche che viene utilizzata come vincolo da imporre all'ensemble canonico. Questo processo di binnaggio prevede di dividere in parti uguali la differenza tra distanza massima e distanza minima, per poi distribuire le distanze intermedie. Nel caso del modello μ, in cui le due gaussiane sono molto distanti e molto strette in rapporto alla distanza reciproca, è ragionevole immaginare che i bin riempiti nel vettore utilizzato come vincolo siano solo quelli alle estremità. Le distanze, infatti, saranno per metà molto grandi (nodo in G_0 - nodo in G_μ), e per tre quarti molto piccole (nodo-nodo nella stessa distribuzione), mentre i bin corrispondenti ai valori intermedi non saranno occupati. Come conseguenza, i valori delle distanze, e quindi delle entropie di singolo nodo, saranno meno distribuite rispetto ad un caso omogeneo in cui tutti i bin di distanza sono occupati.

3.2.4 σ model

Nel secondo modello la distribuzione che governa la posizione spaziale della seconda metà dei nodi è una distribuzione gaussiana in cui è variata sensibilmente la deviazione standard, $G_\sigma = G(\sigma_{var}, \mu_0)$. La posizione della prima metà dei nodi è governata, come nel modello precedente, dalla distribuzione a riposo G_0.

La distribuzione risultante delle entropie di singolo nodo è mostrata, in funzione della connettività, in figura 3.13. In rosso sono mostrate le entropie relative ai nodi

governati da G_0, in giallo le entropie dei nodi governati da G_σ con $\sigma_{var} = 100$.

La notevole separazione delle due distribuzioni mostrata in figura 3.13 è visibile anche in un plot delle entropie di singolo nodo in funzione dell'indice dei nodi. Si osserva infatti la presenza di due distribuzioni diverse governanti le entropie nella prima e nella seconda metà dei nodi.

Un comportamento analogo è esibito nella distribuzione delle differenze ΔS^0 tra la situazione variata σ-model e la situazione spaziale a riposo, mostrata in figura 3.17

Si è analizzata quindi la distribuzione delle differenze punto-punto ΔS^{00} tra le entropie della situazione spaziale variata e le relative entropie nel caso configurazionale, in assenza di una struttura geometrica in cui i nodi sono posizionati. Un'operazione del genere avrebbe come effetto di palesare la variazione del comportamento delle entropie dovuta alla variazione della loro situazione spaziale, eliminando la dipendenza delle S_i dalle specifiche connettività k_i, raddrizzando in tal modo la distribuzione. La distribuzione delle differenze è mostrata in figura 3.19, si vede che effettivamente la dipendenza dalle connettività è stata eliminata, ottenendo una distribuzione raddrizzata. La distribuzione delle ΔS^{00} (figura 3.20) presenta un forte picco nei valori piccoli, dovuto ai nodi governati da G_0, e un comportamento più distribuito per i nodi governati da G_σ

considerazioni Nelle figure 3.13, 3.15, 3.17 si osserva una netta separazione, in termini di distribuzione delle entropie di singolo nodo, tra i nodi appartenenti al primo sottogruppo e quelli appartenenti al secondo. Questa separazione è confermata nell'analisi delle medie e deviazioni standard delle distribuzioni di entropie di singolo nodo, come in tabella 3.2. È evidente che, al contrario di quanto accade nel modello μ, le due distribuzioni non hanno più una configurazione simmetrica, e ciò si riflette in termini di media (più bassa per la distribuzione con σ maggiore) e deviazione standard (più alta) delle entropie di singolo nodo relative.

In termini di confronto con la situazione 0, le entropie dei nodi appartenenti al primo sottogruppo G_0, sono distribuite con una deviazione standard nettamente inferiore. Questo comportamento può essere schematizzato come uno *schiacciamento* dei valori in questione rispetto alla situazione 0, dovuto alla ri-normalizzazione sulla lunghezza massima delle distanze che avviene durante l'imposizione del vincolo sui binnaggi spaziali nella massimizzazione dell'entropia. Questi bin, infatti, rispetto alla situazione zero, apparterranno ad un intervallo massimo-minimo molto più alto, per la presenza di distanze più lunghe dovute alla distribuzione G_σ, e il gruppo di nodi appartenenti a G_0 risulterà di conseguenza in una distribuzione molto piccata di valori piccoli di distanza, che verrà molto probabilmente inclusa in uno o pochi bin. La conseguenza di ciò è che un gruppo di nodi esibirà entropie molto vicine tra loro rispetto al totale, schiacciando la distribuzione nel modo osservato.

Figura 3.13: Valori di S_i nella situazione spaziale variata σ-**model** con $\sigma_{var} = 100 \times \sigma_0$ e $\mu_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_σ

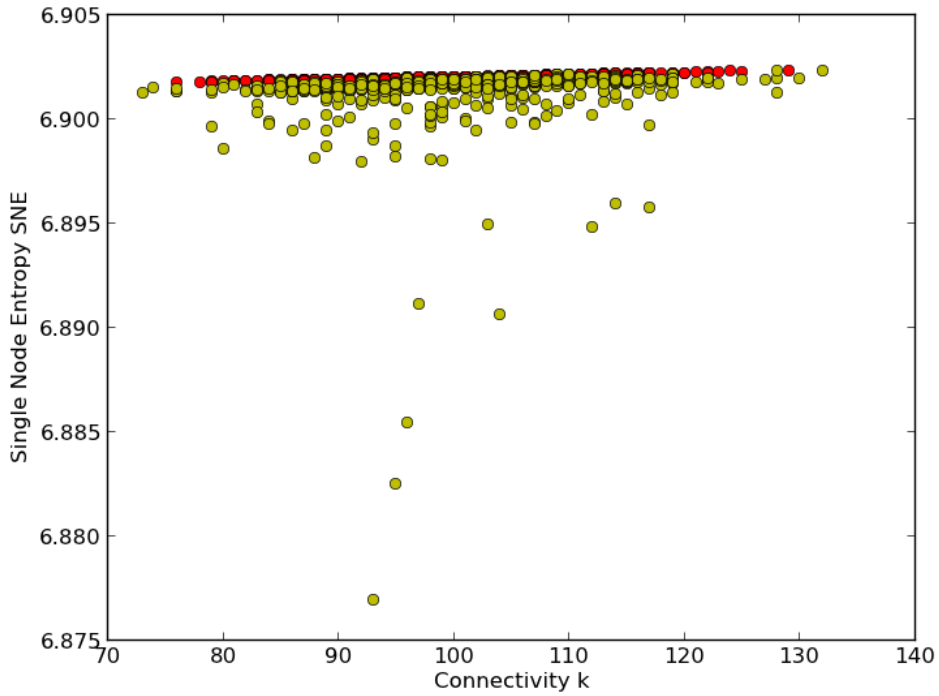


Figura 3.14: uno zoom sulla distribuzione mostra una separazione sensibile dei due sotto-gruppi

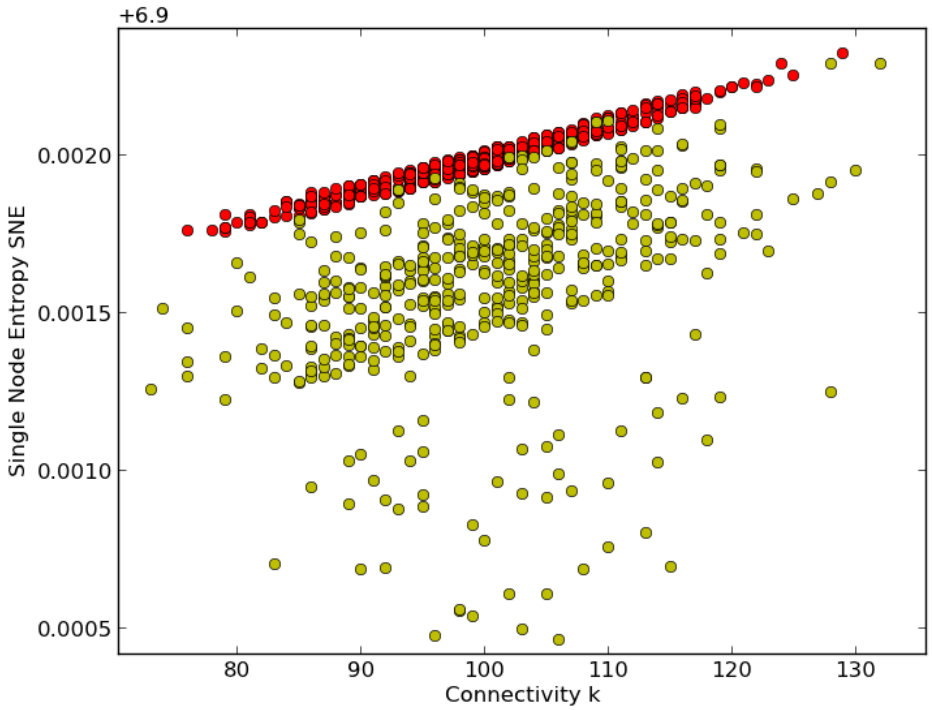


Figura 3.15: Valori di S_i nella situazione spaziale variata **σ-model** con $\sigma_{var} = 100 \times \sigma_0$ e $\mu_0 = 1$ in funzione dell'indice del nodo i nella matrice di adiacenza.

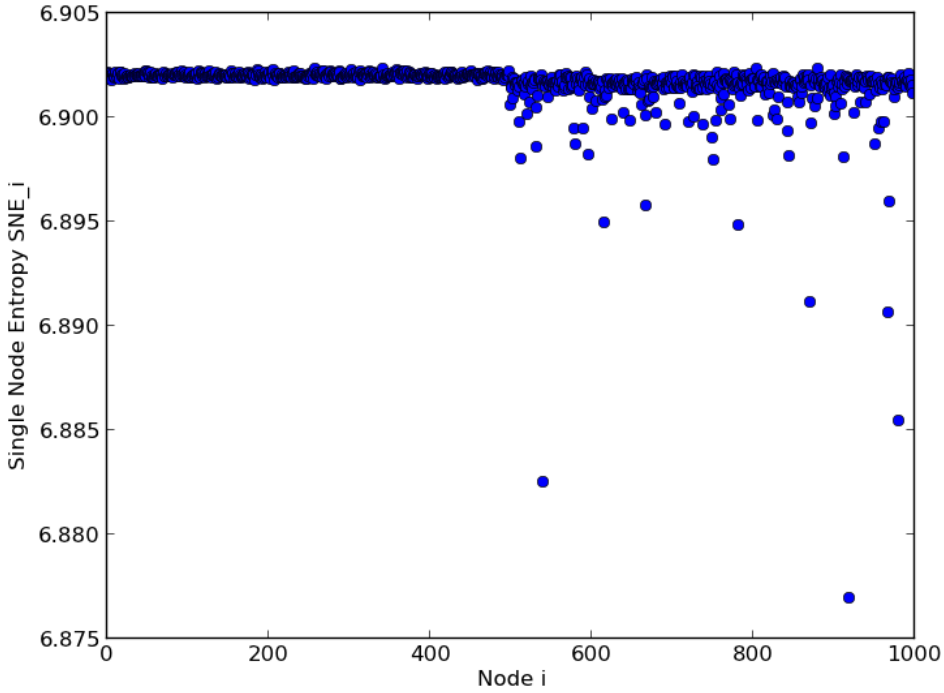


Figura 3.16: uno zoom sulla distribuzione mostra una separazione netta tra le due metà dei nodi, governate da diverse distribuzioni spaziali.

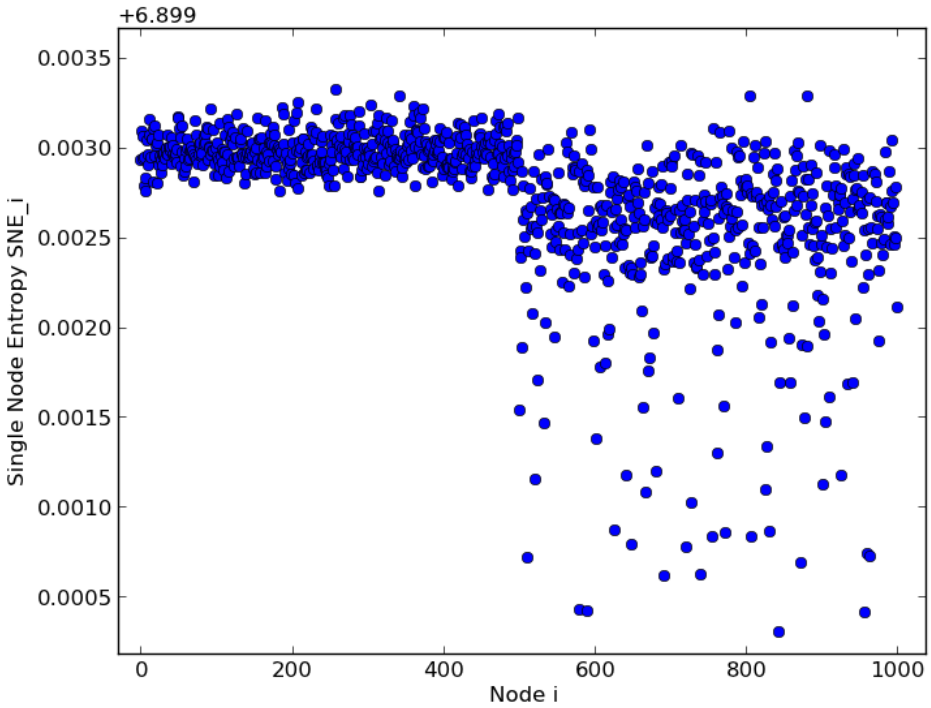


Figura 3.17: Valori di ΔS^0 tra la situazione spaziale variata σ-**model** con $\sigma_{var} = 100$ e la situazione a riposo.

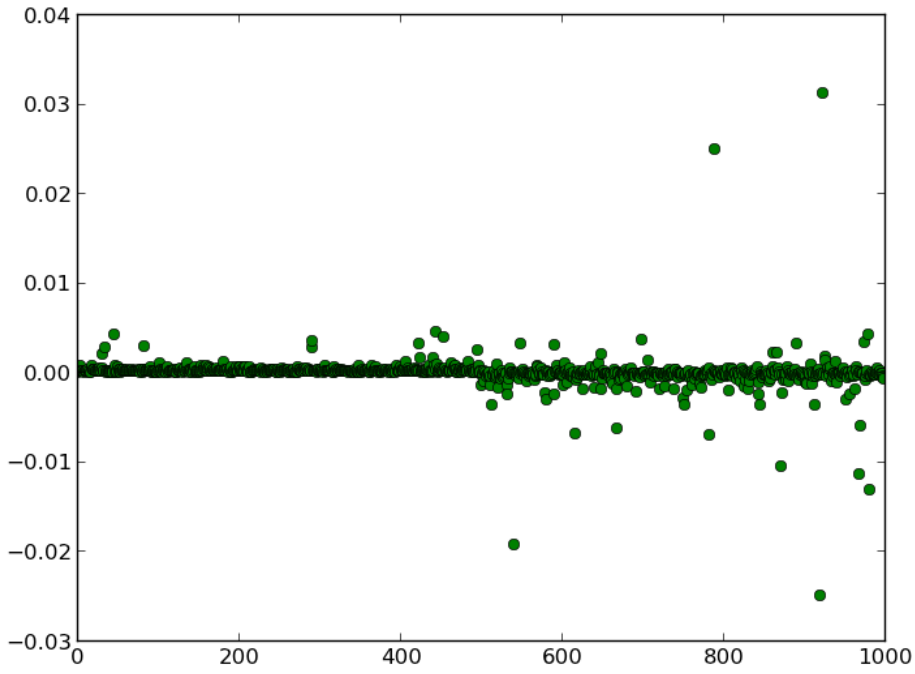


Figura 3.18: Zoom sulla distribuzione

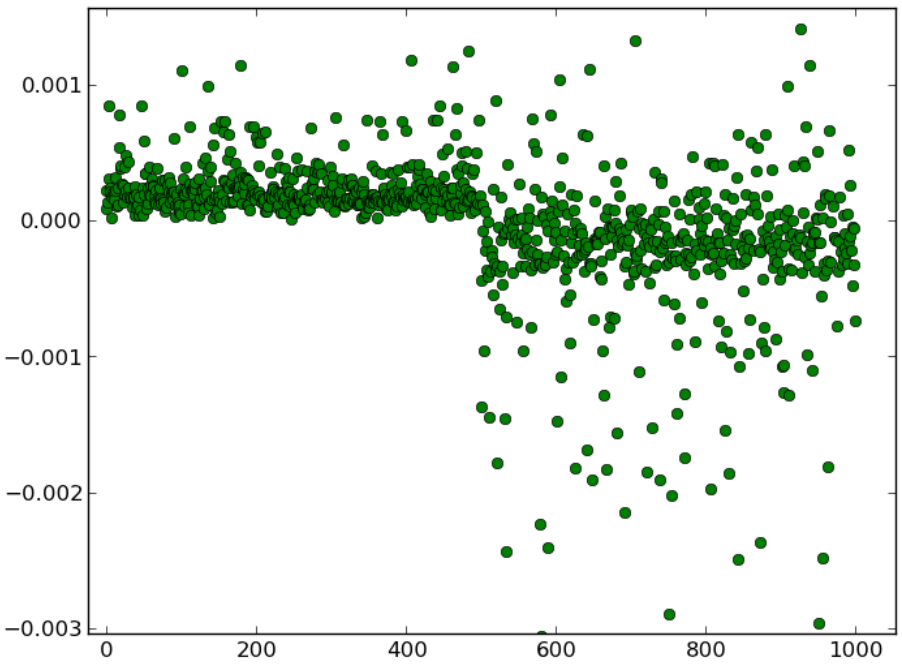


Figura 3.19: Valori di ΔS^{00} tra la situazione spaziale variata σ-**model** con $\sigma_{var} = 100$ e la situazione configurazionale, in funzione della connettività k_i. Non è mostrata dipendenza dalle connettività.

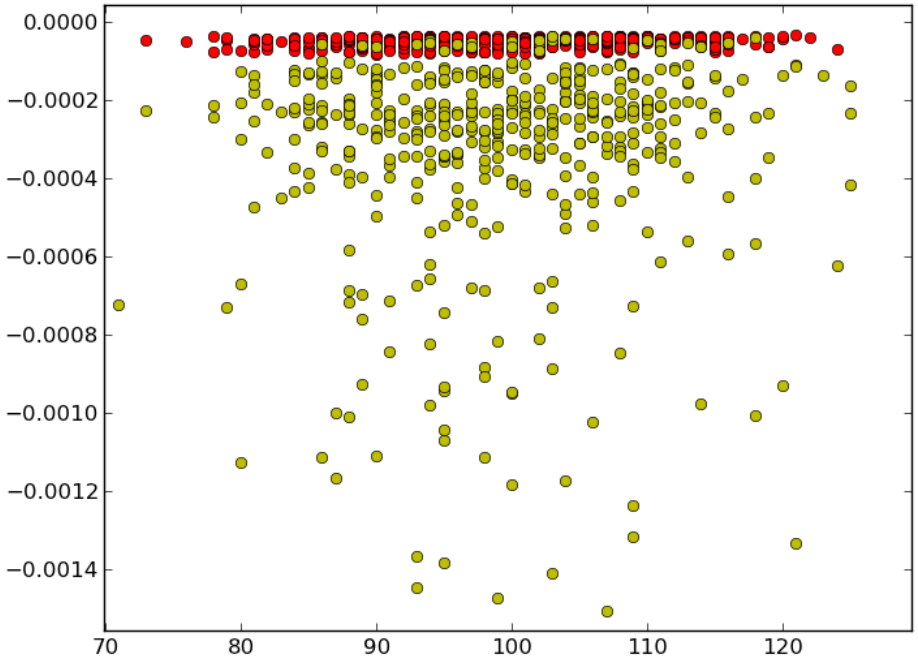


Figura 3.20: Distribuzione di ΔS^{00} all'interno del network.

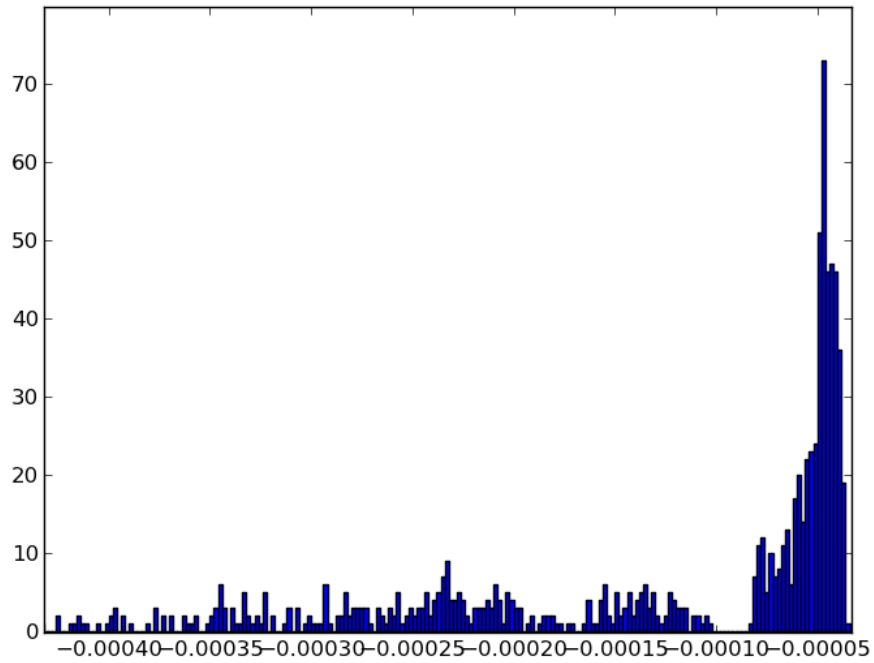


Tabella 3.2: Valori di media deviazione standard per i due gruppi di nodi nel σ**-model**

σ **-model**	media	deviazione standard
globale	6.901597	0.001388
nodi $\in G_0$	6.901984	0.000098
nodi $\in G_\sigma$	6.901211	0.001882
globale 0-sit	6.901600	0.001374

3.2.5 Toy models per l'entropia di singolo nodo S^{TS}

Il procedimento di modellizzazione per l'entropia di Teschendorff-Severini è lo stesso utilizzato nelle sezioni 3.2.3 e 3.2.4: si confrontano due situazioni variate, dove metà dei nodi della matrice delle distanze ha posizioni sottostanti una distribuzione particolare, con una $0 - situation$ in cui tutti i nodi sottostanno alla stessa distribuzione, in cerca di variazioni significative a livello locale e/o globale delle entropie di singolo nodo.

Anche in questi modelli la situazione zero distribuisce le distanze secondo una gaussiana con media e deviazioni standard unitarie. Le entropie di singolo nodo della situazione zero sono mostrate nella figura 3.21:

3.2.6 μ-model S^{TS}

La prima situazione variata è analoga a quella descritta nella sezione 3.2.3: la prima metà dei nodi è distribuita spazialmente secondo una gaussiana $G_0 = G(\mu_0, \sigma_0)$, mentre le posizioni della seconda metà seguono una distribuzione centrata in $\mu_{var} = 100 \cdot \mu_0$, quindi secondo la distribuzione $G_\mu = G(100 \cdot \mu_0, \sigma_0)$.

I risultati del modello sono graficati in figura 3.23, e le caratteristiche delle distribuzioni sono riportate in tabella 3.3.

Tabella 3.3: Valori di media deviazione standard per i due gruppi di nodi nel μ**-model**

μ **-model**	media	deviazione standard
globale	0.861958	0.019067
nodi $\in G_0$	0.861998	0.019147
nodi $\in G_\mu$	0.861918	0.018988
globale *0-sit.*	0.947619	0.011944

considerazioni Dalle figure 3.23 e 3.24 si evince che non c'è una differenza significativa in termini di comportamento esibito dal gruppo variato rispetto al gruppo a riposo. Anche qui, infatti, valgono le considerazioni di simmetria fatte nel modello in sezione 3.2.3, alla luce delle quali è possibile giustificare l'invarianza delle due distribuzioni dal punto di vista di distanze reciproche.

Figura 3.21: Valori di S^{TS} nella situazione spaziale a riposo con $\mu_0 = 1$ e $\sigma_0 = 1$. In rosso i nodi appartenenti alla prima metà, in giallo quelli nella seconda metà. Non si osserva chiaramente differenziazione tra i due gruppi.

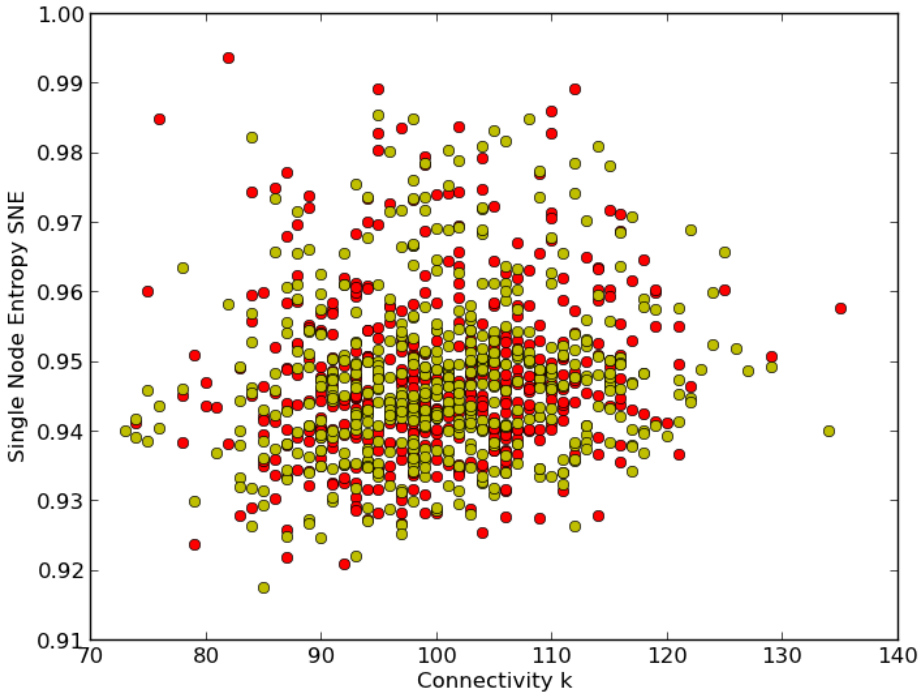


Figura 3.22: Valori di media deviazione standard di S^{TS} per la *0-situation*, ossia tutti i nodi appartenenti alla stessa distribuzione

$0 - situation$	media	deviazione standard
globale	0.947619	0.011944

Figura 3.23: Valori di S^{TS} nella situazione spaziale variata μ-**model** con $\mu_{var} = 100 \times \mu_0$ e $\sigma_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_μ

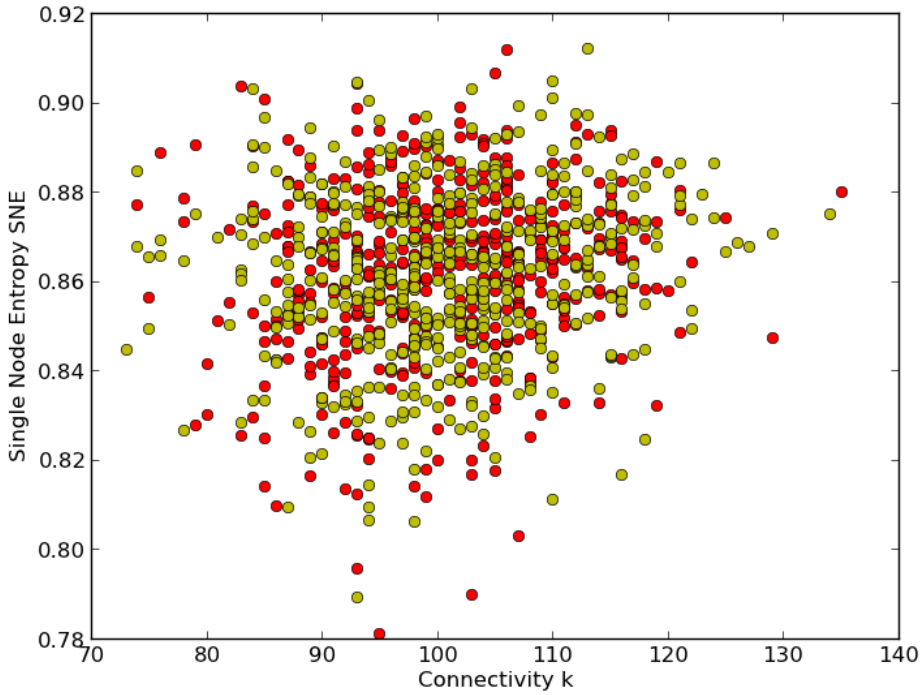


Figura 3.24: Valori di S^{TS} nella situazione spaziale variata μ-**model** con $\mu_{var} = 100 \times \mu_0$ e $\sigma_0 = 1$, distribuiti in funzione dell'indice del nodo i

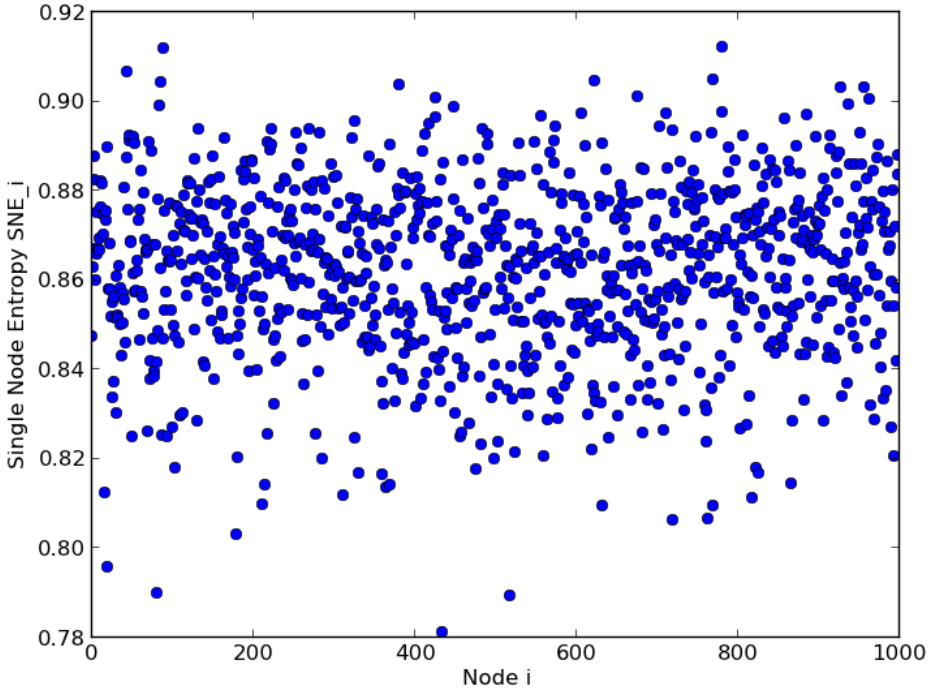


Tabella 3.4: Valori di media deviazione standard per i due gruppi di nodi nel **σ-model**

σ **-model**	media	deviazione standard
globale	0.876704	0.078020
nodi $\in G_0$	0.805694	0.021101
nodi $\in G_\sigma$	0.947714	0.040549
globale $0 - sit.$	0.947620	0.011944

A differenza del modello 3.2.3, tuttavia, si osserva una diminuzione sensibile dell'entropia in termini di media globale. Questo è giustificabile dal punto di vista della teoria dell'informazione: avendo un maggior numero di distribuzioni che governano le posizioni spaziali, e di consegienza le p_{ij}, un quantitativo maggiore di informazione è contenuto nella distribuzione globale delle probabilità, diminuendone in un certo senso l'ignoranza che l'osservatore ha nei suoi confronti, e quindi l'entropia di Shannon. In termini qualitativi, infatti, possiamo prevedere che i nodi appartenenti a due distribuzioni diverse (interpretata nel formalismo di Teschendorff-Severini come correlazione positiva), avranno valori molto alti di p_{ij}, mentre quelli appartenenti alla stessa distribuzione avranno un valore molto più basso.

Mancando tutti i valori intermedi di distanze, e considerato soprattutto che metà delle probabilità sono molto piccole rispetto all'altra metà, la distribuzione delle p_{ij} è molto *ordinata* rispetto alla distribuzione generale gaussiana della $0 - situation$, ed è naturale quindi immaginare un valore di entropia di Shannon minore rispetto a quest'ultima.

3.2.7 σ-model S^{TS}

Il modello visto nella sezione 3.2.4 è stato utilizzato anche per caratterizzare la misura di Teschendorff-Severini. La situazione zero è la stessa vista nei modelli precedenti, ossia tutti i nodi $\in G_0$, mentre la situazione spaziale variata prevede che metà dei nodi siano distribuiti secondo una distribuzione gaussiana con la deviazione standard sensibilmente aumentata, ossia $G_\sigma = G(\mu_0, 100 \cdot \sigma_0)$.

L'andamento della misura e i valori statistici del caso sono riportati in figura 3.25 e 3.26 e in tabella 3.4

considerazioni Dal comportamento mostrato in figura 3.25 e 3.26 è evidente una separazione netta tra le distribuzioni delle entropie di singolo nodo relative ad un sottogruppo rispetto ad un altro. In particolare si nota che il secondo sotto-gruppo, governato dalla distribuzione G_σ ha un comportamento molto simile alla situazione a riposo, anche in termini di media e deviazione standard, mentre il primo sottogruppo, governato da G_0, presenta un valore medio sensibilmente inferiore rispetto alla situazione a riposo.

Figura 3.25: Valori di S^{TS} nella situazione spaziale variata σ-**model** con $\sigma_{var} = 100 \cdot \sigma_0$ e $\mu_0 = 1$, in funzione della connettività k. In rosso i nodi a riposo, in giallo quelli governati dalla distribuzione variata G_σ

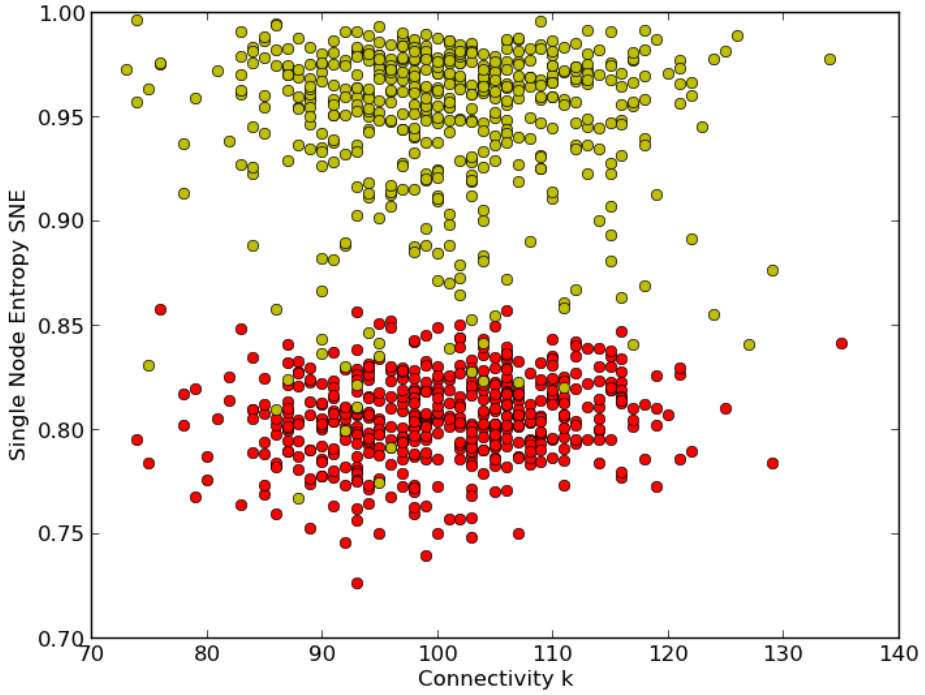
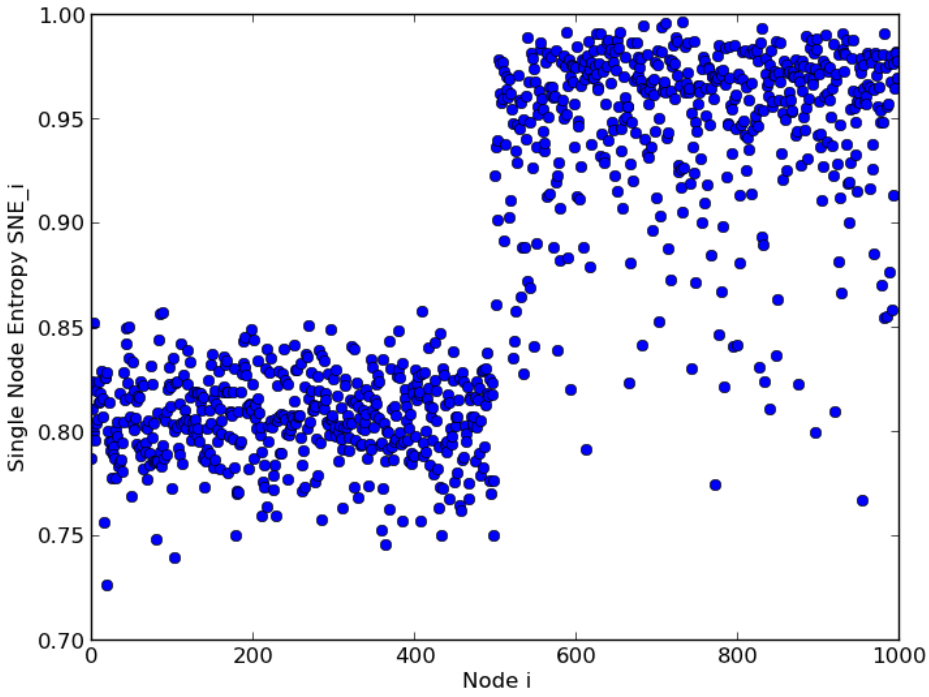


Figura 3.26: Valori di S^{TS} nella situazione spaziale variata σ-**model** con $\sigma_{var} = 100 \times \sigma_0$ e $\mu_0 = 1$ in funzione dell'indice del nodo i nella matrice di adiacenza



Questo comportamento è spiegabile ancora una volta in termini informazionali. Nel procedimento di passaggio dai valori di distanza relativa C_{ij} alle p_{ij}, infatti, la distribuzione delle C_{ij} viene normalizzata sulla somma totale in modo da sommare a uno. In questo modo si appianano, di fatto, due distribuzioni di uguale forma ma con valori diversi, come G_0 e G_σ: entrambe gaussiane centrate sullo stesso valore, ma con una deviazione standard molto diversa che risulta tuttavia riscalata al momento della normalizzazione.

Al contrario, i nodi che appartengono alla distribuzione G_0 nel contesto spaziale variato, ossia in presenza di metà nodi governati da G_σ, rappresentano un elemento di *ordine* nella distribuzione totale, avendo di fatto metà delle p_{ij} risultanti dai collegamenti interni alla distribuzione G_0, e quindi molto piccoli e molto vicini tra loro rispetto ai valori delle distanze tra loro e gli altri nodi, ampiamente distribuiti, appartenenti alla gaussiana G_σ.

L'entropia di Shannon della relativa distribuzione delle p_{ij}, composta quindi da metà valori molto piccoli e simili tra loro e metà valori distribuiti gaussianamente, sarà sensibilmente più bassa di quella di una distribuzione semplicemente gaussiana, caso della distribuzione nella $0 - situation$.

Questo caso è molto simile a quello visto nella sezione 3.2.6 come organizzazione delle p_{ij} a entropia inferiore.

Capitolo 4

Conclusioni

In questo lavoro è stata introdotta e caratterizzata una misura di entropia per un singolo nodo, S_i, nella prospettiva di distinguere, all'interno di un network, nodi aventi differenti caratteristiche. Nello stesso contesto è stato studiato il comportamento di una seconda misura di entropia di singolo nodo proveniente dalla letteratura, qui chiamata S^{TS} , ai fini di ottenere un confronto critico tra le due in termini di performance nel distinguere i nodi appartenenti a due sottogruppi diversi e di comportamento al variare delle caratteristiche proprie di questi sottogruppi.

caratterizzazione analitica È stata fatta una prima caratterizzazione analitica della distribuzione delle entropie di singolo nodo, secondo la misura qui definita, all'interno dell'ensemble configurazionale. Si è quindi misurata l'andamento della media delle entropie di singolo nodo in funzione della connettività media, ottenendo una funzione costante per valori non piccoli.

Si è ottenuto che, attraverso alcune approssimazioni, è possibile ottenere un'espressione analitica dell'entropia di singolo nodo in funzione della connettività del nodo in questione, normalizzata, come $S_i \propto k'_i \log k'_i$. È stato possibile giustificare successivamente, con l'approssimazione di campo medio, l'indipendenza della media di queste entropie dalla connettività media del network, ottenendo una forma per questa misura come $\langle S_i \rangle \simeq \left(1 - \frac{1}{N}\right) \log N$, espressione che si è vista rappresentare inoltre una buona previsione per i valori misurati.

Il confronto tra le due misure è stato effettuato nel contesto dello **spatial ensemble**, ossia dove i nodi del network sono posizionati su una struttura geometrica e si utilizza, nel calcolo della matrice di probabilità p_{ij}, la matrice delle distanze reciproche tra i nodi.

comportamento entropico Si è voluto inizialmente fare un'analisi del comportamento delle misure al variare della randomness della matrice di distanze utilizzata nel calcolo della matrice di probabilità dell'ensemble di appartenenza del network. Secondo l'interpretazione classica dell'entropia come disordine della distribuzione, questa dovreb-

be essere una funzione crescente della randomness, in quanto la distribuzione random rappresenta il massimo disordine possibile.

Nel caso della misura qui definita, la matrice di distanza non è utilizzata direttamente nel calcolo dell'entropia di singolo nodo, ma solo come vincolo nella massimizzazione dell'entropia, la quale risulta nella distribuzione di probabilità p_{ij}. Questo fa sì che il comportamento entropico non sia direttamente correlato alla randomness della matrice delle distanze, e il risultato osservato conferma un'invarianza dell'entropia rispetto questo parametro.

Per la misura S^{TS}, si è ottenuto invece un comportamento inverso. Questo è giustificabile dal fatto che la definizione di Teschendorff-Severini utilizza la matrice delle distanze direttamente nel calcolo dell'entropia, considerando quindi le singole distanze come probabilità p_{ij}, dopo opportune normalizzazioni. Con questo metodo il massimo ordine, distanze tutte uguali, corrisponde alla distribuzione ad entropia massima, in quanto si ha la distribuzione delle probabilità sulla quale si possono fare meno assunzioni, ossia quella in cui tutte le probabilità degli eventi in questione sono uguali, analoga al caso del lancio di una moneta non truccata o di un dado.

toy models Si sono quindi implementati dei modelli in cui la seconda metà dei nodi nel network in questione viene posta in una situazione spaziale variata, al fine di ottenere una fitness per le due misure in termini di capacità di distinguere l'appartenenza dei singoli nodi ai due sotto-gruppi.

Al fine di ottenere una migliore separazione dei gruppi, nel caso della misura S_i qui introdotta, si è voluta eliminare la dipendenza delle entropie dalla struttura configurazionale del network in modo da palesare eventuali differenze dovute alle posizioni spaziali.

Per far ciò sono state analizzate le differenze ΔS^0, tra la situazione spaziale variata e una situazione *a riposo* dove tutti i nodi seguono la stessa distribuzione, e ΔS^{00} tra la situazione spaziale variata e le entropie configurazionali, ossia ottenute senza l'imposizione della matrice delle distanze.

μ model Nel primo modello le posizioni dei nodi sono state distribuite, rispettivamente nel primo e nel secondo gruppo, secondo una gaussiana $G(\mu, \sigma)$, con $\mu = \sigma = 1$ e secondo una gaussiana variata $G(100 \cdot \mu, \sigma)$.

Per l'entropia S_i non si è osservata una variazione significativa dalla distribuzione nel caso a riposo, dove tutti i nodi sono disposti secondo $G(\mu, \sigma)$.

Per l'entropia S^{TS} si è osservata una variazione globale della disposizione, ma nessuna informazione significativa in grado di distinguere l'appartenenza dei nodi al primo o al secondo gruppo.

σ model Nel secondo modello la posizione spaziale del secondo gruppo di nodi è stata disposta secondo una gaussiana con diversa deviazione standard, ossia $G(\mu, 100 \cdot \sigma)$, mentre la prima metà è stata disposta secondo la distribuzione a riposo $G(\mu, \sigma)$.

In questo caso si è osservata, per entrambe le misure, una netta separazione tra i nodi appartenenti al primo sottogruppo e quelli appartenenti al secondo. Questa separazione è confermata nell'analisi delle medie e deviazioni standard delle distribuzioni di entropie di singolo nodo.

Nel caso dell'entropia S_i si osserva un compattamento delle entropie appartenenti alla distribuzione a riposo, verificato in termini di deviazione standard del sottogruppo che risulta decisamente diminuita. Le entropie relative al primo sotto-gruppo sono inoltre tendenzialmente più alte di quelle del sotto-gruppo variato. Non si osserva quindi una situazione bi-modale, ma una situazione in cui un gruppo è ampiamente distribuito e l'altro è più compatto.

Nel caso dell'entropia di Teschendorff-Severini, la misura è egualmente in grado di distinguere l'appartenenza ai sotto-gruppi dei singoli nodi, ma presenta delle differenze con l'entropia di singolo nodo qui introdotta. Le entropie sono infatti disposte in forma bi-modale, e le due distribuzioni dei sotto-gruppi sono simili in termini di varianza. Inoltre, mentre la misura S_i risulta minore per i nodi regolati dalla distribuzione variata, S^{TS} risulta invece maggiore. I risultati sono giustificabili direttamente nel contesto della teoria dell'informazione, in quanto i valori della matrice di distanza sono utilizzati direttamente come p_{ij} nel calcolo delle entropie.

In conclusione, entrambe le misure, con le dovute precisazioni di contesto e differenze reciproche, si sono mostrate in grado di distinguere l'appartenenza dei nodi ai sotto-gruppi caratterizzati da una disposizione spaziale differente, nel caso in cui le differenze tra le distribuzioni spaziali siano in termini di varianza. È auspicabile l'applicazione, in futuro, della misura qui definita a dati reali, ad esempio nel contesto della transcrittomica, per verificare che costituiscano osservabili in grado di distinguere le proprietà dei singoli geni, o proteine, in diverse situazioni ad alto livello a confronto.

Bibliografia

[1] Ginestra Bianconi. *A statistical mechanics approach for scale-free networks and finite-scale networks* Chaos 17, 026114 (2007)

[2] Ginestra Bianconi. *The entropy of randomized network ensembles* Europhys. Lett. 81, 28005 (2008).

[3] Ginestra Bianconi. *Entropy of a network ensemble.* Physical Review E 79, 036114 (2009).

[4] P. Erdős, A. Rényi. *On the evolution of random graphs.* Publ. Math. Inst. Hung. Acad. Sci 5, 17 (1960)

[5] E. T. Jaynes (1957) *Information theory and statistical mechanics* Physical Review 106:620

[6] G.Menichetti, G.Bianconi, E.Giampieri, G.Castellani e D.Remondini. *Network Entropy measures applied to different systemic perturbations of cell basal state.*

[7] Juyong Park, M. E. J. Newman. *The statistical mechanics of networks* Phys. Rev. E 70, 066117.

[8] C.E. Shannon *A Mathematical Theory of Communication* Bell System Technical Journal, vol. 27, pp. 379-423, 623-656 (July, October 1948)

[9] Andrew E Teschendorff, Simone Severini *Increased entropy of signal transduction in the cancer metastasis phenotype* BMC Systems Biology, 4:104 (2010)

[10] Joy A. Thomas and Thomas M. Cover, *Elements of Information Theory* John Wiley & Sons, (2006)

[11] M. Tribus, E.C. McIrvine *Energy and information* Scientific American, 224 (September 1971).

[12] G. Turchetti *Meccanica, sitemi dinamici, complessità dagli atomi agli automi, il passaggio al vivente e forse al pensante* Paesaggi della complessità e inseparabilità dei mondi. Ed. Luciano Boi, MIMESIS (2010)

[13] `http://en.wikipedia.org/wiki/History_of_entropy`